

└ Statistical Tools for Analysis

- Basic concepts
- Graphical analysis and data exploration
- Statistical design for watershed studies

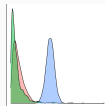
This is going to be a whirlwind tour of some statistical tools available to analyze watershed data. I know everyone is coming in here with different levels of knowledge of statistics and some (or many) dislike stats. So I'm trying not to get too deep, but provide a roadmap for decision-making and point you to additional resources as needed. If you want to become really familiar with the concepts I'll point you to the Practical stats course the we have provided with Dr. Helsel in the past (note sure if/when it will continue with his retirement) but again the USGS statistical methods in water resources book is a must have reference for anyone doing statistical analysis of water quality data.

└ Base concepts

- Statistical distributions
- Measures of central tendency
- Concentrations vs loads

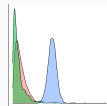
I am going to go over three different different concepts so we are all on generally the same page. First we will talk about statistical distributions because they serve an important role in how we determine what types of data analysis we can do. Then a couple measures of central tendency that we can use to describe data, and final a short discussion about concentrations and loads.

└ Statistical distributions



Let's talk about statistical distributions, does anyone have a good definition or practical explanation for what a distribution is?

└ Statistical distributions

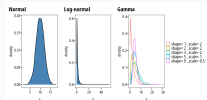


A statistical distribution is a rule or function that describes the probability that a variable takes on some range of values.

A simple definition is that a statistical distribution is a rule or mathematical function that describes the probability that a variable takes on some range of values.

└ Statistical distributions

- For the most part we deal with normally distributed, log-normal, or gamma distributions.
- Influences our choice of statistical tests.

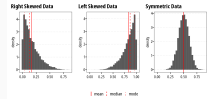


The shape and the area under the normal distribution is mathematically described with two parameters, the mean and standard deviation. The log-normal distribution is normally distributed when values are log transformed, and described by the log mean and the log standard deviation. The shape and area of the Gamma distribution is described using two parameters names the shape and rate parameters. The Gamma distribution is always positive and skewed, it can appear similar to the log-normal distribution. For topics we cover today you really only need to be aware that we can use parametric tests on normally distributed data, and rely on nonparametric tests for other distributions.

Measures of central tendency

Measures of central tendency

- **Mean:** Sum divided by number of samples.
- **Median:** Midpoint of all values or mean of two middle values.
- **Mode:** Most likely value.



Measures of central tendency, or averages, represent a central or typical value from a sample. The most common are mean, medians and modes. For symmetric datasets, the mean median and mode are roughly the same. For skewed datasets, the mean gets pull towards the tail or more extreme values. We typically use median in skewed datasets.

└ Measures of central tendency

The **geometric mean** is typically used with data that are extremely variable (bacteria).

- **Geometric mean:** Average of log transformed values converted back to real (base 10) number.
- Calculate by exponentiating the mean of log transformed values:
- OR, the n th root of the product of n numbers.

The other commonly used measure is the geometric mean. We typically use this with extremely skewed data such as bacteria. It is calculated by log transforming the values then taking the average, then exponentiating it back.

Measures of central tendency

Here is an example using some E. coli data. We see the extreme values pull the Mean much higher than the median and geometric mean.

Example Dataset

Sample number	E. coli
1	5
2	26
3	50
4	30
5	890
6	15
7	100

- Mean = 159
- Geomean = 43
- Median = 30

- The geomean is the average but less influenced by the few extreme values.
- Median still represents the middle value.

Does anybody want to describe the differences between water quality loads and concentrations? Why might we be interested in one or the other?

STATISTICAL TOOLS FOR ANALYSIS

2024-03-03

└ Concentrations and Loads

Concentrations represent the amount of pollutant at a given point in time.

- Instantaneous effect
- Density based units, mg/L, etc/100ml, etc.



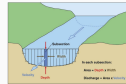
Concentrations are the amount of substance dissolved or suspended in a volume of liquid. This is measured with a grab sample, lab analysis or sensor and represents a specific point in time. Concentration is constantly variable across space and time. We grab a snapshot of it.

└ Concentrations and Loads

Loads represent a cumulative amount of pollutant

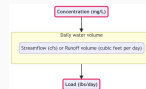
Loads represent mass over time

- Cumulative effect
- Units are mass based, pounds/year, kg/month, etc.



Concentrations and Loads

Water volume over time is required to convert concentration to loads.



In order to convert from concentration to load we have to have the the total volume of water in a given day or whichever time unit you are interested in. Typically we use mean daily flow from a streamflow gage or the total volume reported by a runoff monitoring station.

Concentrations and Loads

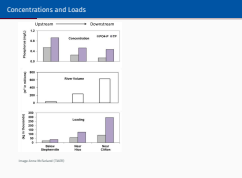
Convert concentration to loads:

$$\text{Load}(\text{lbs}/\text{day}) = \frac{\text{mg}}{\text{L}} \times 28.32 \frac{\text{L}}{\text{ft}^3} \times \frac{\text{ft}^3}{\text{day}} \times 1\text{E6} \frac{\text{mg}}{\text{kg}} \times 2.2046 \frac{\text{lbs}}{\text{kg}}$$

• Always check units and conversion factors.

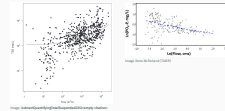
This is an example calculation going from mg/L concentration to pounds per day. First we convert the concentration in liters to concentration in cubic feet. We multiply that by the volume of water in cubic feet to get the total volume in milligrams. Some simple conversions are available to convert from milligrams to pounds.

Concentrations and Loads



Here we see streamflow or discharge volume can effect loads and concentrations differently. The top plots show that as we move upstream to downstream phosphorus concentration decline with increasing river volume. However loads increase with river volume. What does this tell us? River volume is the primary driver of loading. Think about the units we use, relatively large variations in concentration have small impacts on loads, but small variations in flow will make big differences in total loads.

└ Concentrations and Loads



This doesn't mean streamflow is not important to consider for concentrations. Within one site, streamflow can drive variation in pollutant concentration. For NONpoint sources, concentrations may increase, while point source dominated pollutants may decrease with increased flows.

└ Concentrations and Loads

These are different measurements, why are we concerned about both?

- To estimate progress in rivers/streams we want to understand changes in concentration.
- To estimate progress in the watershed we want to understand changes in flow-normalized loads.
- To estimate progress in estuaries or lake/reservoir ecosystems we want to understand total loads.

Robert Winch (USGS)

So why all the bother with loads and concentrations? Typically, if we are interested in stream/river health or exposure and risk for humans, we want to understand how pollutant concentrations are responding. Exposure risks or ecosystem responses are typically based on some concentration thresholds. If we are interested in how watershed water quality is changing across the landscape we have to assess loads because we can't measure concentration at every point across the landscape. Better yet, we use flow-normalization methods to account for variations in precipitation and runoff and allow us to compare loads between wet and dry periods. Finally if our primary objective is an estuary or lake system, we are primarily interested in changes in the total loads delivered to the system. Assuming the volume of water is relatively constant over time in a lake or estuary, the total mass of pollutants delivered provides an appropriate metric.

Before we move into exploratory data analysis, any questions?

└ Exploratory Data Analysis

First step in any data analysis is to plot your data.

- Graphical methods provide quick visual summaries of data.
- Easily interpreted.
- Describes essential information more easily than numbers alone.

Your first task with any assessment is to plot your data. There are a plethora of tools available to us now days to do this quickly and easily. We do this to get a quick and easy to intpret understanding of the data. Summary numbers alone can cause us to miss important information that are plainly obvious when plotted.

└ Exploratory Data Analysis

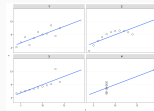


Figure 1: Four different datasets with the same mean, variance, correlation, slope and intercept. Dataset is known as Anscombe's quartet.

This figure is often used to emphasize the importance of plotting your data. These 4 sets of data are called Anscombe's quartet. If you calculate the mean, variance, correlation, slope and intercept of these data without looking, you would make some incorrect statements about the underlying data.

└ Available graphical methods

- Histograms and density plots
- Quantile plots (cumulative density function)
- Boxplots
- Probability plots
- Scatterplots

These are your basic graphical tools to explore data, routines for most of these are available in Excel and all major statistical software. These tools will help you assess the distribution, variance, mean or median, and general relationship between variables.

Histograms and density plots

- Histograms plot the count of observed values within equally spaced bins.
- Displays the distribution, skewness, and variability of the data.
- Density plots are smoothed versions of histograms.

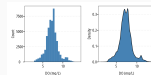


Figure 2: Histogram and density plot of 15-minute DO measurements.

On the left, is an example of a histogram, the height of each bar is equal to the count of observations within that bin of values. Each bin is equally spaced. The histogram shows if your data is skewed and the relative variability of your data. One potential issue with histograms is choosing the bin size. You may need to make a couple of different histograms increasing and decreasing the number of bins or the size of the bins. The density plot on the right is basically a smoothed version of the histogram. Density plots reduce the need to manipulate bin size, many statistical software include automatic smoothness selection methods for these plots, but you have the option of manually adjusting the bandwidth of the underlying density estimator used to produce these plots.

Quantile plots

Quantile plots

- Provides information about the distribution of observed values.
- Shows the probability that a random variable will be less than or equal to specific value x .
- Also called empirical cumulative distribution functions (ecdf).
- A flow duration curve is an inverse version of the ecdf using descending ranks instead of ascending ranks.

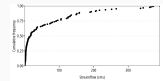


Figure 3: Quantile plot of 2-years of mean daily streamflow values

Quantile plots are approximations of the cumulative distribution function, or the probability that a variable will be less than or equal to some value. The vertical axis are quantiles 0 to 1 representing the smallest and largest possible values. The quantiles are calculated by ranking the values and applying a plotting position formula. Quantiles such as 0.5 or the median can be quickly identified, once you get used to them, the skew and distribution can be identified. Big advantages are that all the data is displayed, there is no interpretation of bins/categories/or smoothing functions. The disadvantage of the quantile plot is that they are clearly harder to interpret when you are not familiar with them.

Boxplots

Boxplots

Boxplots are concise displays of the median, variation, skew, and outliers. These can also be used to compare attributes between datasets or sites.

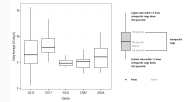


Figure 4: Boxplots of dissolved oxygen concentrations at 5 sites.

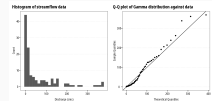
Boxplots are a concise display of various summary statistics and distribution. The centerline represents the median, the height of the box is the variance or spread, and you can infer the skew based on the size of the boxes above or below the median. Outliers are shown as points beyond the boxplot whiskers. Most statistical software has functions to generate boxplots. One disadvantage of boxplots is they don't visualize the data points so you might miss details like bimodal distributions. This can be overcome by overlaying points on the boxplot itself.

└ Probability plots

Probability plots also called q-q plots plot quantiles estimated from the data against quantiles from any theoretical distribution, such as the normal, log-normal, or gamma distributions. The histogram on the left shows a skewed dataset that might be log-normal or gamma distributed. The Q-Q plot confirms the sample quantiles generally follow the quantiles from the theoretical Gamma distribution quite well.

Probability plots

Also called a quantile-quantile (Q-Q) plot. This is the quantile plot generated earlier plotted against quantiles from a theoretical distribution. These are used to evaluate how well the data fits against distributions such as the normal, log-normal, or gamma distribution.



Onto statistical design. My goal is to provide a road map for choosing or applying appropriate statistical methods without diving into theory. There are a couple of free text references at the end of this presentation that go much more in depth. But the primary goal is to understand what data you have, your study design, then choose the right statistical approach. Which, spoiler, there isn't always just one correct way to do things.

└ Single watershed study



- Comparing concentrations before and after implementation.
- Parametric t-test or nonparametric Rank-sum test.
- Null hypothesis: average concentrations before and after implementation are equal.
- Flow-weighting averages can be used to emphasize high flows.

In a single watershed study we can compare the means before and after implementation. Typically we will use a parametric t-test on log-transformed concentrations, or the non-parametric Rank Sum test on raw values. If you have flow or discharge data, then we can use something called flow-weighted averages.

└ Flow-Weighted Average

Here is an example of a flow-weighted average. You multiply the observed concentration by the measured flow and divide by the sum of weights. The weighting factor, flow, gives more emphasis to concentrations at high flow than at low flows.

Example:

Concentration (mg/L)	Flow (cfs)
0.45	10
2.30	0.01
0.75	15

$$\frac{(0.45 \times 10) + (2.30 \times 0.01) + (0.75 \times 15)}{10 + 0.01 + 15} = 0.63 \text{ mg/L}$$

$$\frac{0.45 + 2.30 + 0.75}{3} = 1.17 \text{ mg/L}$$

└ Permutation Test

- **Parametric tests** on log-transformed values provide information about the geometric mean.
- **Non-parametric tests** tell you about the median.
- If you need information about the mean use a **permutation test**.

Typically we will use a parametric t-test on log-transformed concentrations, or the non-parametric Rank Sum test on raw values. But because we transform log values in the parametric test, we are actually inferring about the geometric mean when we back transform our data after the test. Conversely, non-parametric tests typically internally transform data into ranks, the result is we are detecting differences in medians. Typically this is fine, but make sure you report your results accordingly. If you must report differences in means, we can use permutation tests. Permutation tests are nice because we don't need to know anything about the distribution and it is still valid.

└ Permutation Test

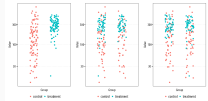
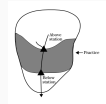


Figure 6: Permutations randomly shuffle that data between groups. Assuming there is no difference between the groups, new reshuffles will have approximately the same differences.

Hopefully this provides a little insight into how a permutation test works. First we calculate our test statistic using the observed data between our two groups. If there is no difference between the two groups if we shuffle the values between them, we should not get a significantly different result. So we shuffle and recalculate the test result. Then re do that 1000 or more times. This gives us a distribution of possible null values of the test statistic to compare our actual results against. The percentage of values that our test statistic exceeds out of all the recalculated values is our p-value. This approach can be adapted to most hypothesis tests which makes it suitable for most watershed study approaches.

└ Single watershed study

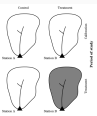


- Compare upstream and downstream concentrations.
- Parametric paired t-test or non-parametric signed rank test.
- Include before and after implementation as factors.
- Parametric Two factor ANOVA or nonparametric Brunner-Dette-Munk test.
- Permutation tests available for each of these.

For a single watershed study with above design, we can use a paired t-test or signed rank test. Preferably you want to include before and after implementation data, this will be setup as a two-factor ANOVA or non-parametric Brunner-Dette-Munk test. The results will let you infer if the downstream water quality after implementation is significantly different than before implementation and/or from the upstream station.

└ Paired Watershed Study

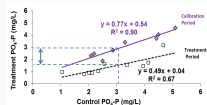
Paired Watershed Study



- Two-factor ANOVA, nonparametric Brunner-Della-Murk test
- Can be setup as a linear regression model
- Treated = $\beta_0 + \beta_1(\text{Control}) + \epsilon$

We can take a similar approach with the paired watershed study. However I'd recommend a linear regression approach because it is flexible enough to incorporate additional variables if needed, like if you want to control for streamflow or other covariates. There are several ways to set it up, but the simplest is to model your response variable, treatment water quality concentration, as a factor of control watershed water quality. If your measurements are not daily, you can use weekly or monthly flow weight averages to pair up data from each watershed in your linear model.

└ Paired Watershed Study



Here is an example from the folks at TIAER. The linear regression in purple is the calibration period and the dashed line is the treatment period. The vertical distance between the lines is the relative reduction in concentration by the BMP treatment. The difference in slope also shows a larger reduction at high runoff concentration values compared to lower runoff concentration values.

└ Paired Watershed Study

Weakness:

- Assumes relationships in water quality between two watersheds.
- Regression and ANOVA approach have parametric assumptions.

Other available tools include: Generalized linear models and generalized additive models which are **semi parametric** statistical tools.

There are a few potential issues with linear regressions in paired watershed studies. First we are assuming that water quality mechanisms in both watersheds are the same, to meet this assumption, both watersheds should be nearby each other with similar land cover and land use. Second this approach requires we meet parametric assumption, such as normally distributed data. To accommodate different types of data, we start to get into more advanced statistical methods, such as generalized linear or generalized additive models.

Cheet sheet

Design	Parametric	Nonparametric	Permutation
Pre/Post	t-test	Rank-sum test	Test sample
Upstream/Downstream	Paired t-test	Signed-rank test	permutation test
Randomized			Permut per-mutation test
Random stratified	Two-Sample (ANOVA or MANOVA)	Discriminant Analysis	Test factor
	Linear regression	test on generalized linear model	permutation test

Here is a short reference sheet to assist with choosing methods based on study design and the type of data you are working with.

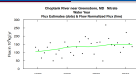
Trend Analysis

Type	Transformed for variable X	Adjusted for variable X
Parametric	Regression of Y on T	Regression of Y on T and X
Nonparametric	Median smoothed trend test	Median smoothed trend test residuals then linear regression of T on X

- Trends can be temporal or spatial.
- T can represent time or distance.
- X can represent streamflow, or precipitation.

With trend tests, we are interested in the relative fluctuation or increase and decrease in concentrations or trends either temporally or spatially. With parametric data, this is simple, use a linear regression modelling response variable (concentration or load) as a function of time or space. Either date transformed to a numeric variable or temporal variable like river mile. We can incorporate seasonality by adding additional terms like day of the year. We often want to adjust for terms like streamflow which can account for the overwhelming amount of variation in water quality. For non parametric approaches, we use the Mann-Kendall test. If we want to incorporate a term like streamflow, we need to first fit a regression model to the flow-water quality relationship and obtain the residuals, or the difference between the expected and observed water quality values. The Mann-Kendall test is then applied to the residuals to obtain the flow-adjusted trend.

Trend Analysis



The USGS developed Weighted Regressions on Time Discharge and Season (WRTDS) tool provides functions for assessing trends in concentration, loads, and **flow-normalized** loads.

- Provides ability to assess loads as if streamflow was consistent from year to year.
- Incorporates non-linear or smoothed trends.
- See [HirschlerGuideExploration2015](#) empty citation.

More recently USGS has developed a tool called WRTDS or weighted regressions on time discharge and season. This statistical tool available in R, allows us to assess actual trends, and flow normalized trends over time. The flow normalized trends can be considered the expected value if streamflow was consistent from year to year. This is an advanced and powerful tool for assessing estimated concentration and loading trends over time. However it does require substantial data, probably a minimum of 100-200 samples over 10-20 years minimum. This figure shows the actual annual nitrate load in a river over 30 years as the dots. You see substantial variation due to changes in streamflow. The green line is the flow normalized load that shows a clear steady increasing trend in load that would occur if streamflows were consistent year to year.

└ Overview

- Plot/explore data.
- Data distribution and assumptions should match the statistical approach.
- Use flow-adjustment or flow normalization if it matches your objective/question.
- Rarely a single correct approach.

So in summary, if nothing else, plot your data multiple ways. Make sure your statistical approach matches your data. Keep your study objective in mind, your approach should match the objective and you might have to take extra steps such as flow adjustments to appropriately answer those study questions. Finally, there is rarely a single approach, we often have to weigh different approaches and recognize specific compromises when making a decision.