STATISTICAL TOOLS FOR ANALYSIS

Fundamentals of Developing a Water Quality Monitoring Plan

Michael Schramm

2024-03-05

TWRI, Texas A&M AgriLife Research

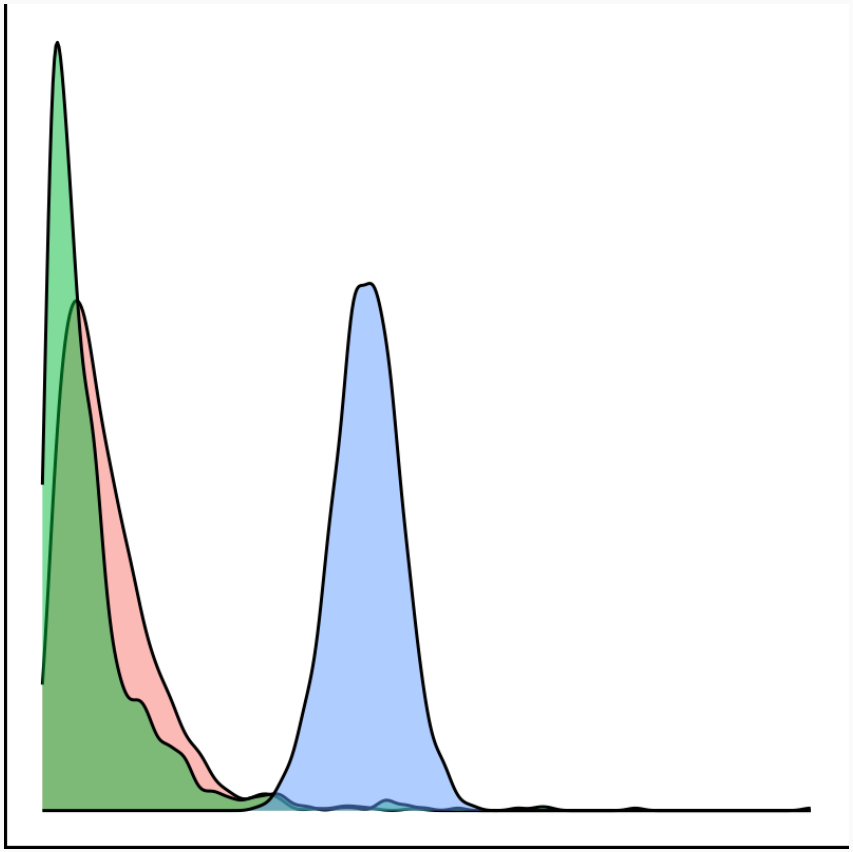
Statistical Tools for Analysis

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

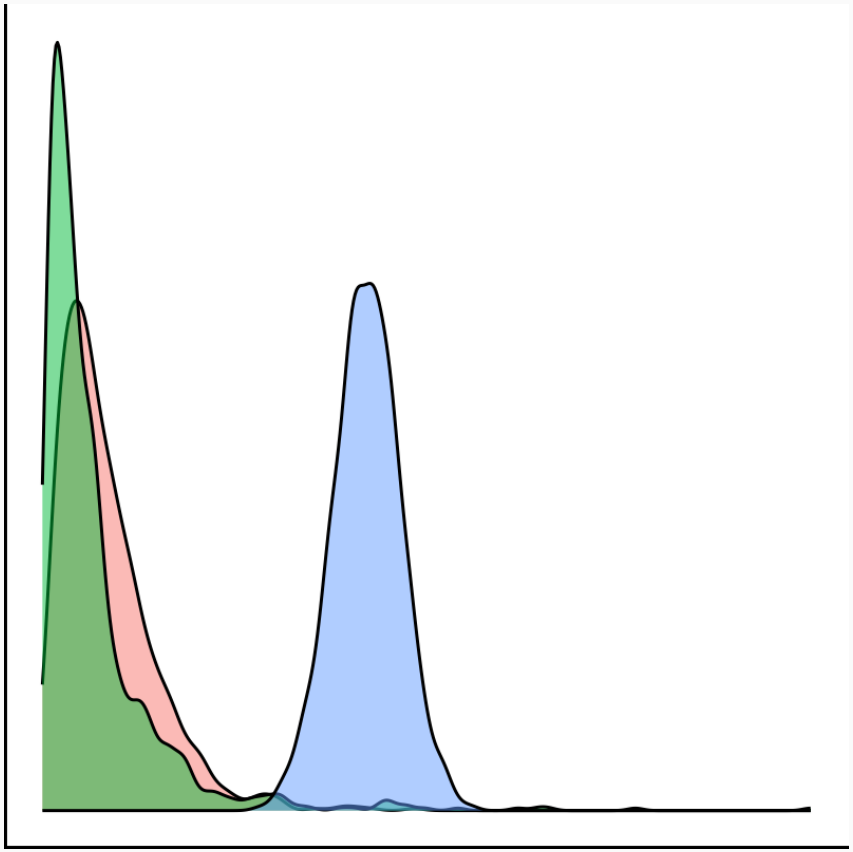
Base concepts

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

Statistical distributions



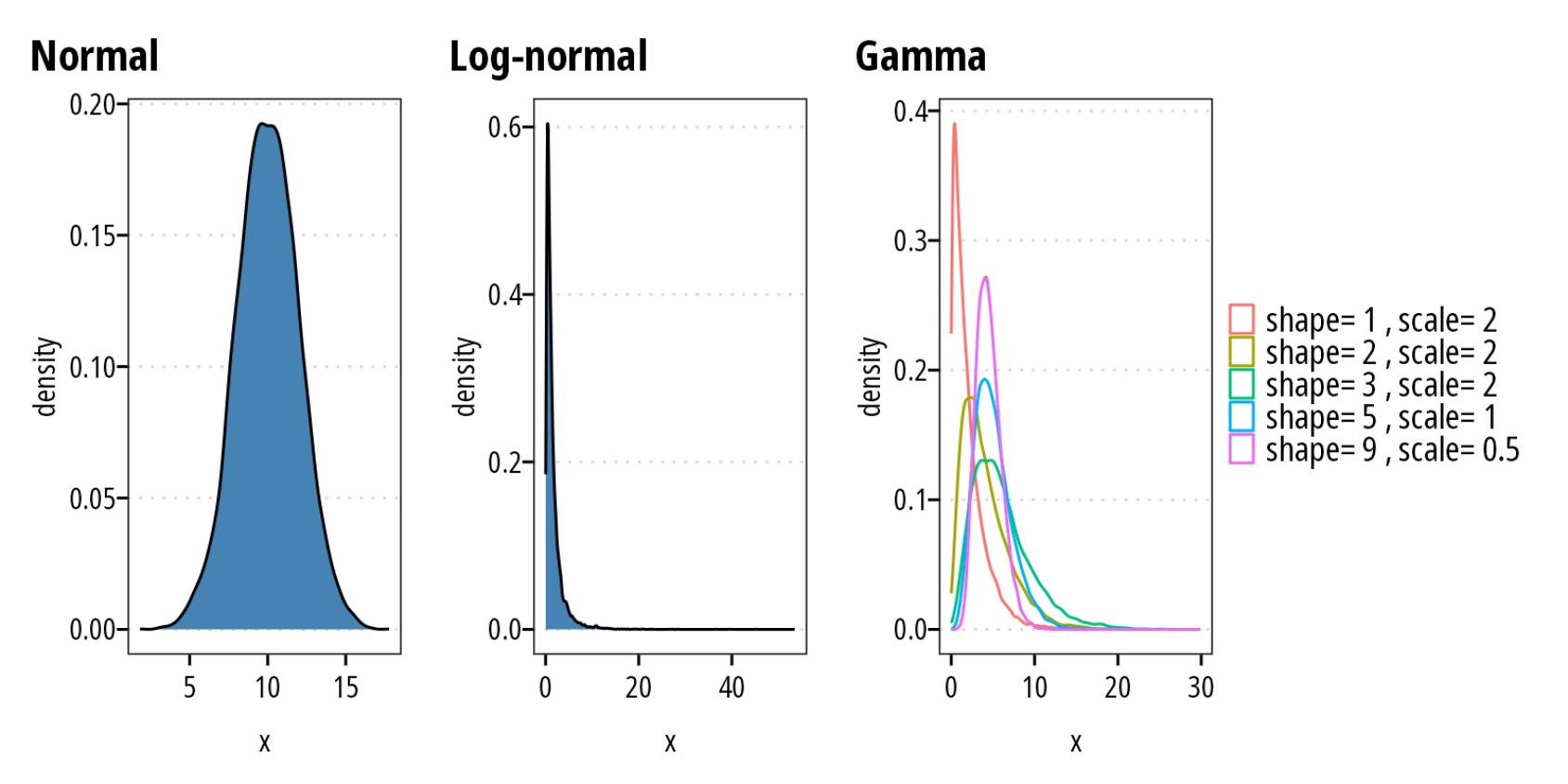
Statistical distributions



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

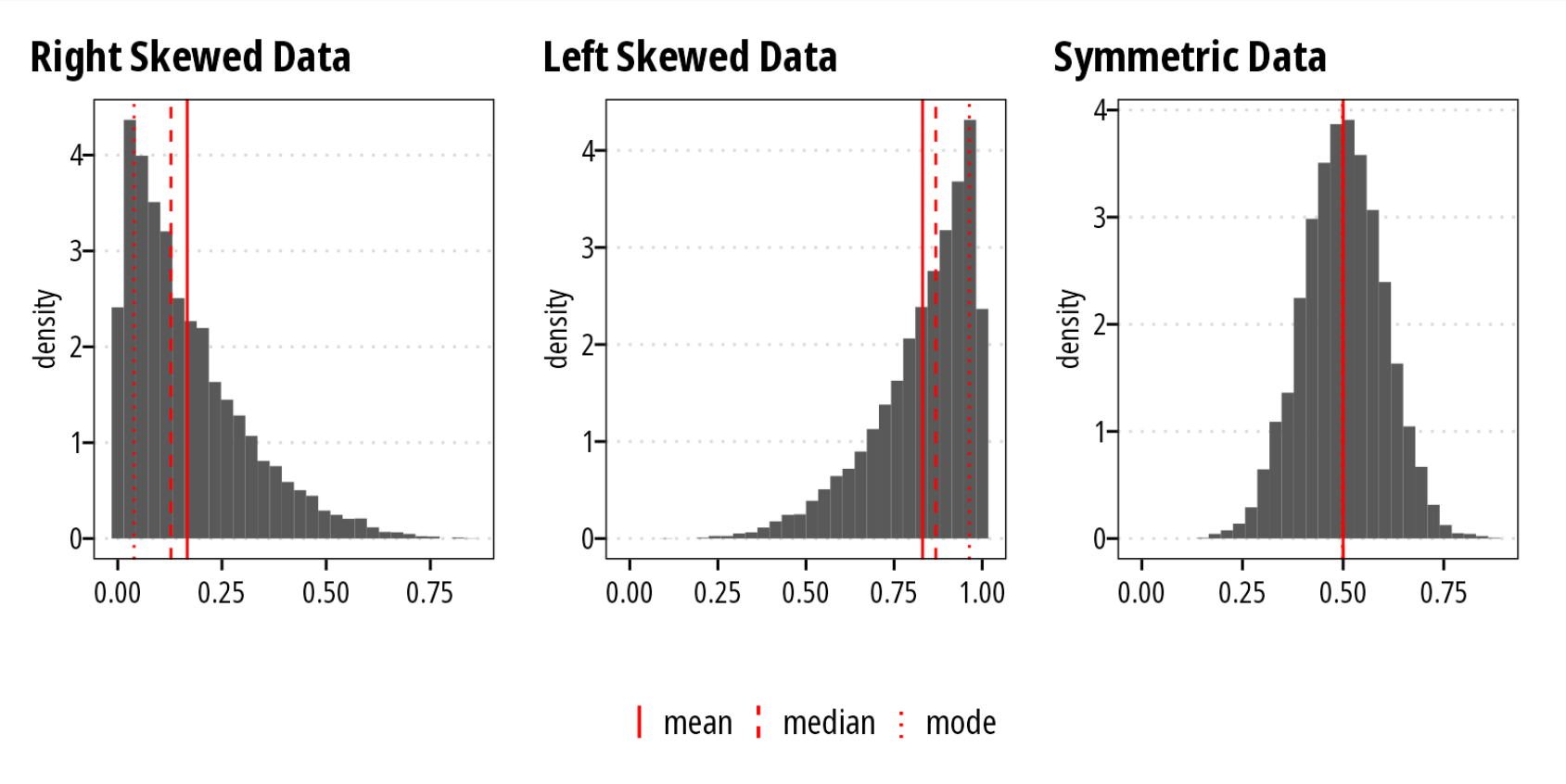
Statistical distributions

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



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

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 nth root of the product of n numbers.

Measures of central tendency

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.

Concentrations and Loads

Concentrations and Loads

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

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

Concentrations and Loads

Loads represent mass over time

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

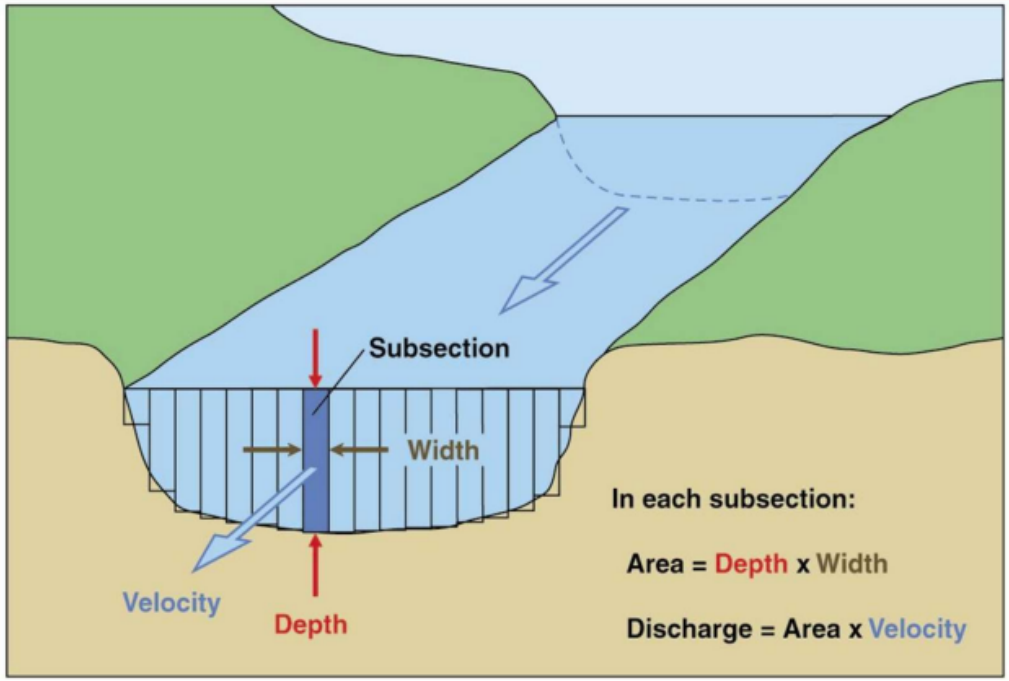
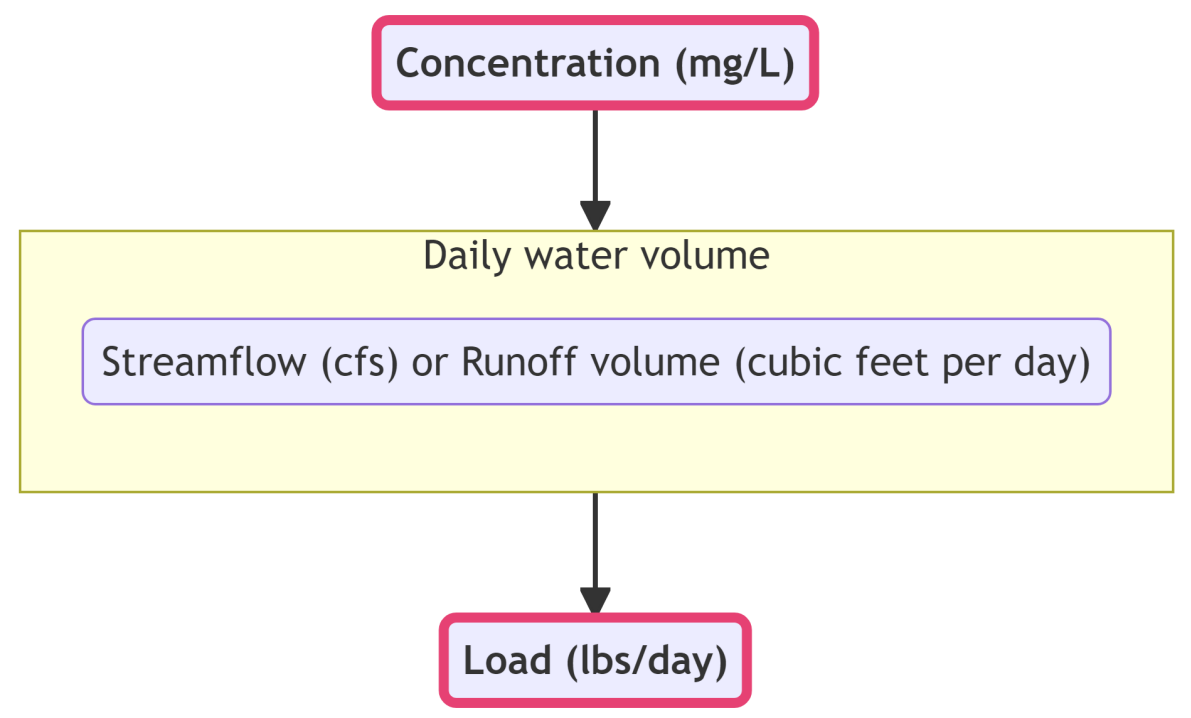


Image USGS public domain.

Concentrations and Loads

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



Concentrations and Loads

Convert concentration to loads:

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

- Always check units and conversion factors.

Concentrations and Loads

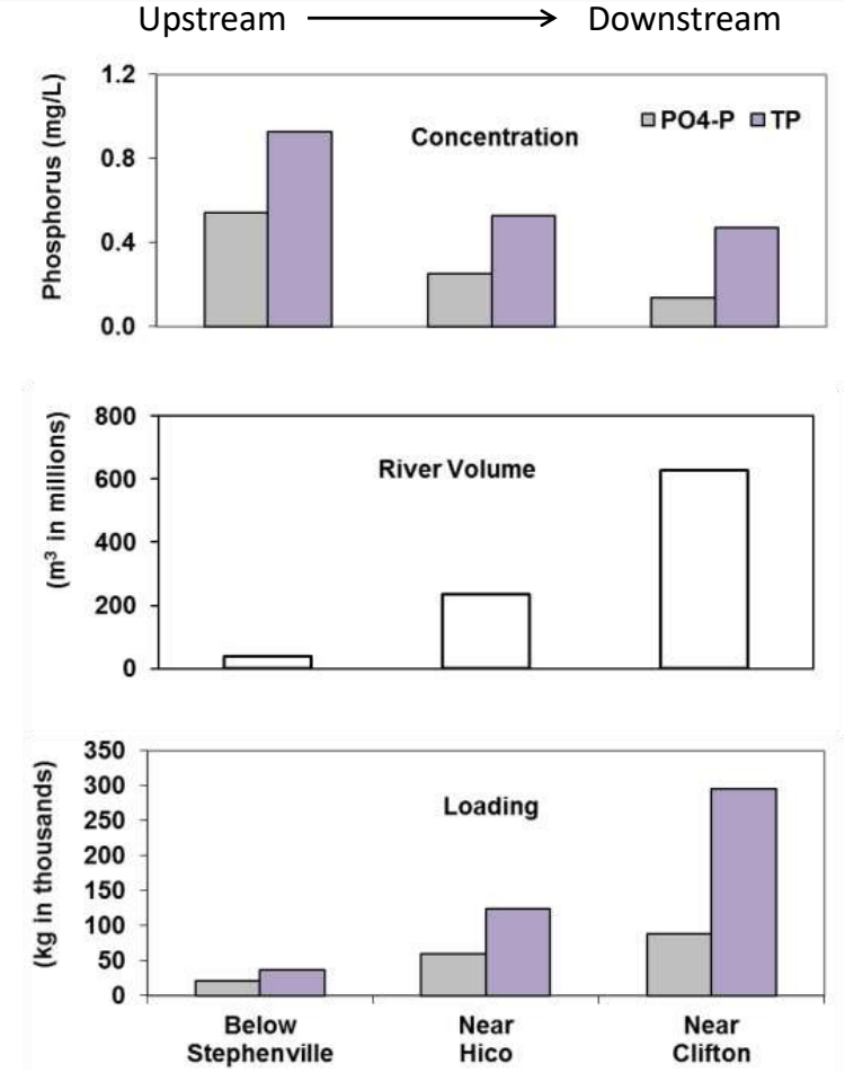


Image Anne McFarland (TIAER)

Concentrations and Loads

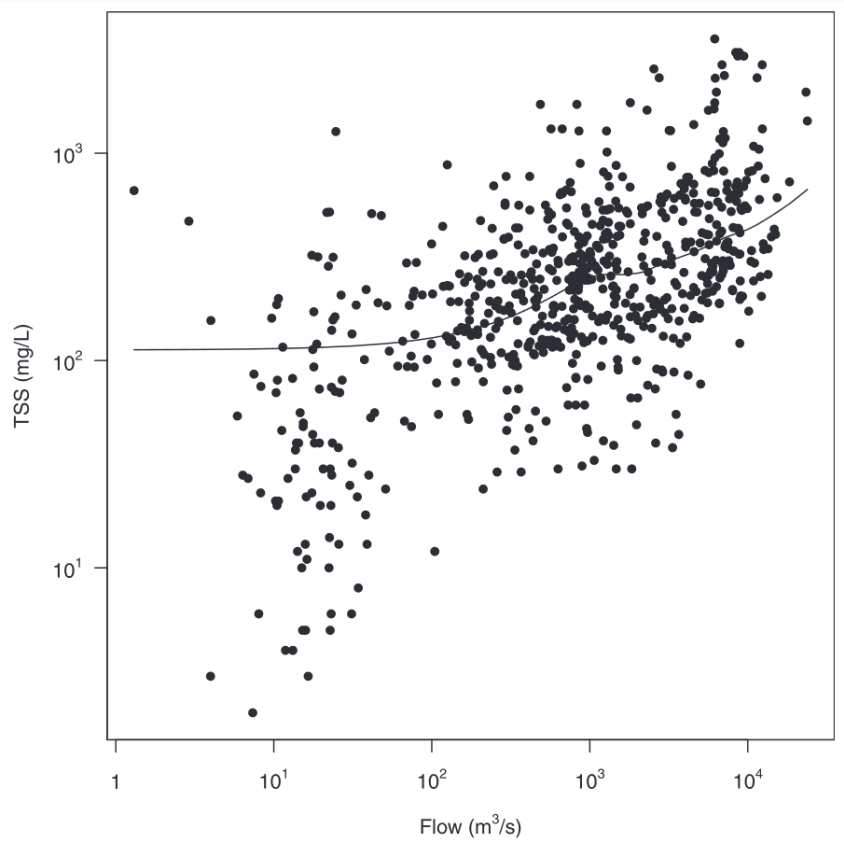


Image: Kuhnert et al. (2012)

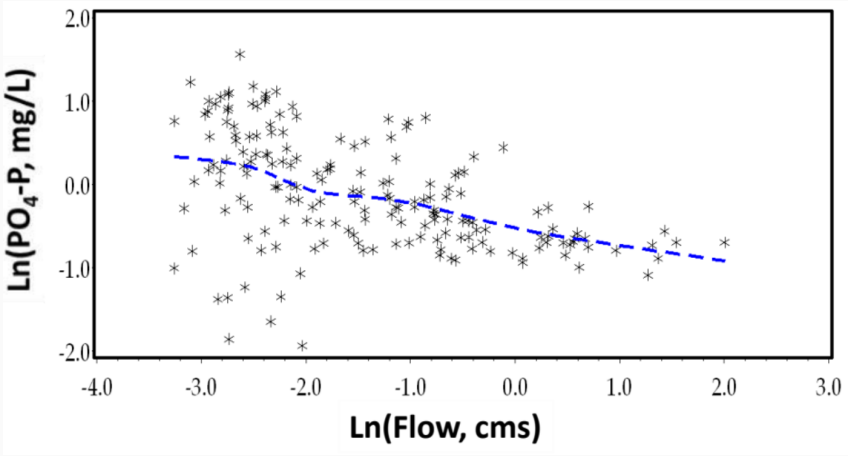


Image Anne McFarland (TIAER)

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 Hirsch (USGS).

Exploratory Data Analysis

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.

Exploratory Data Analysis

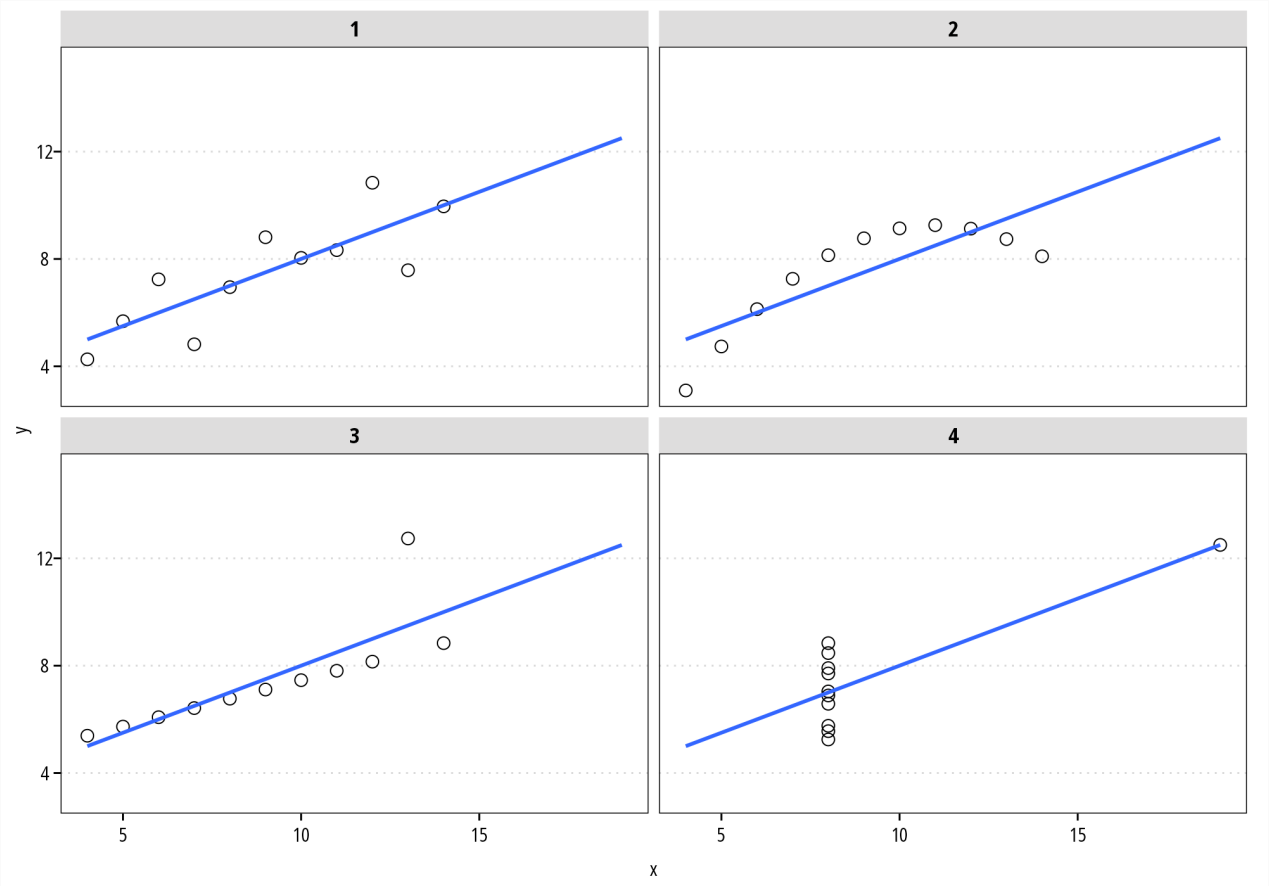


Figure 1: Four different datasets with the same mean, variance, correlation, slope and intercept. Dataset is known as Anscombe's quartet.

Available graphical methods

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

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.

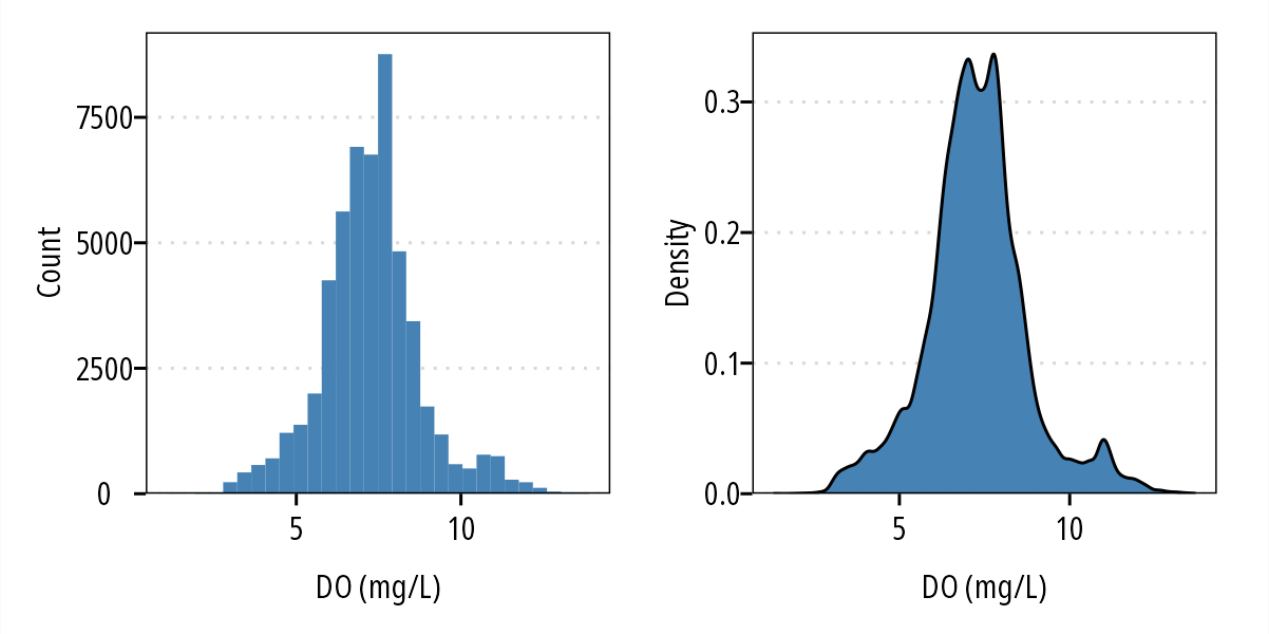


Figure 2: Histogram and density plot of 15-minute DO measurements.

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.

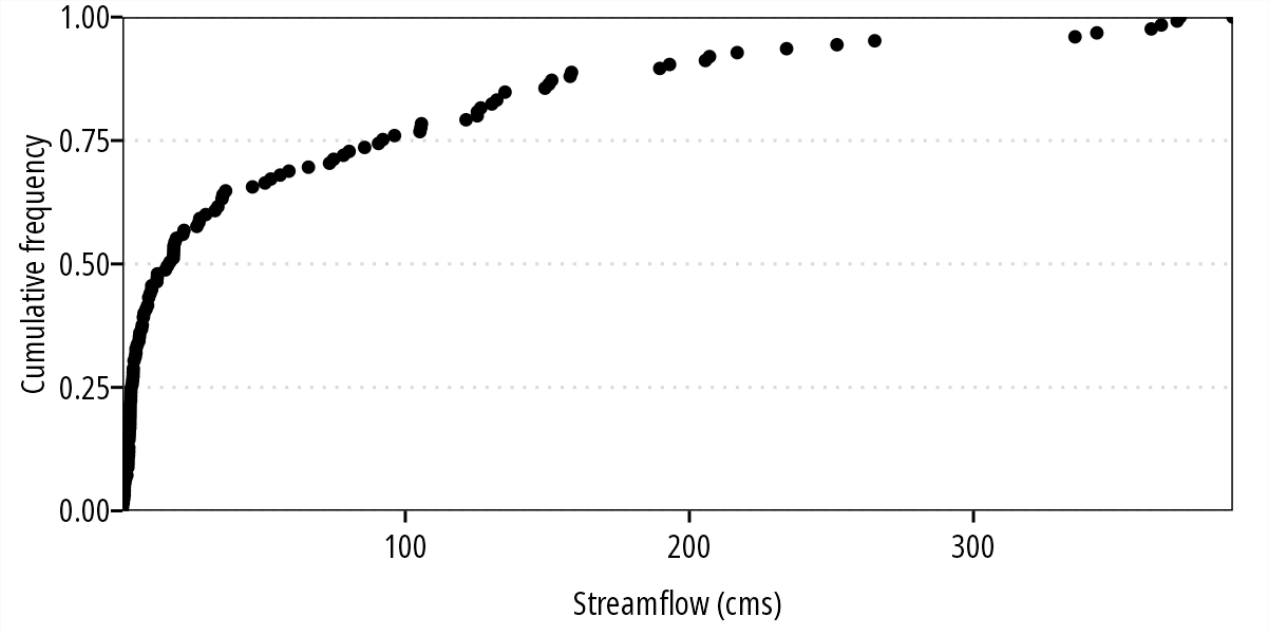


Figure 3: Quantile plot of 2-years of mean daily streamflow values

Boxplots

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

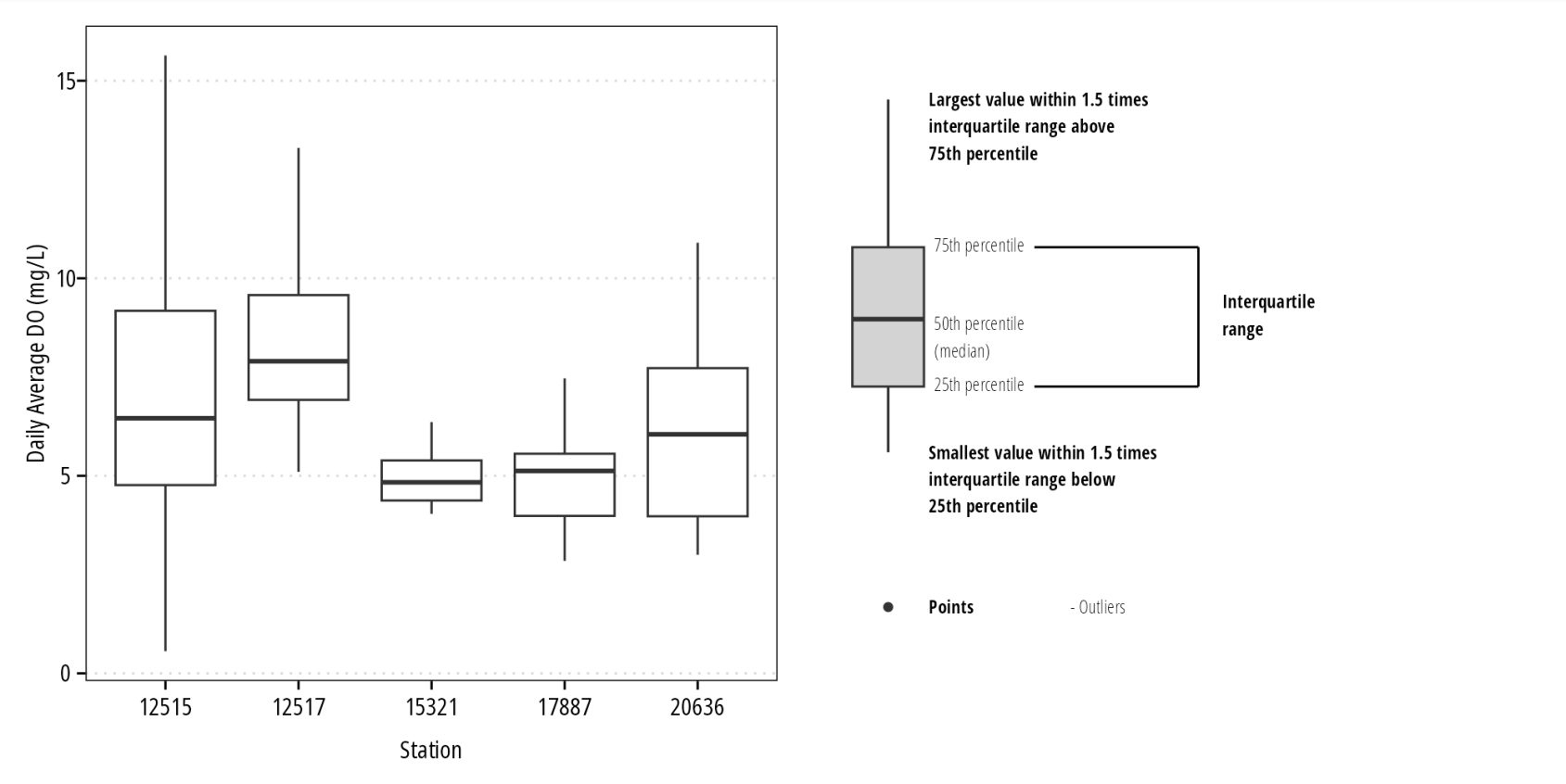
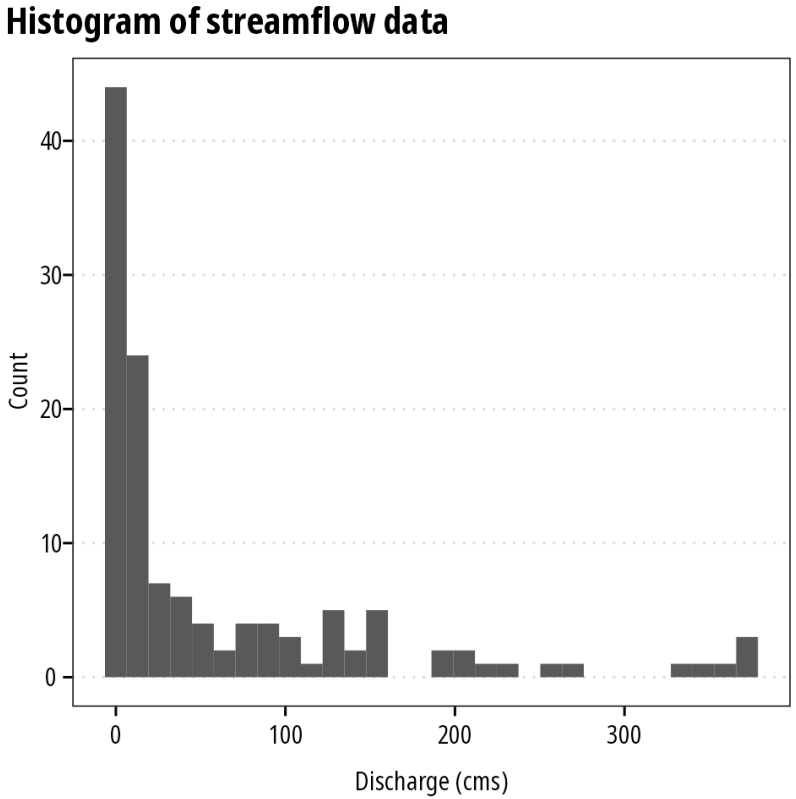
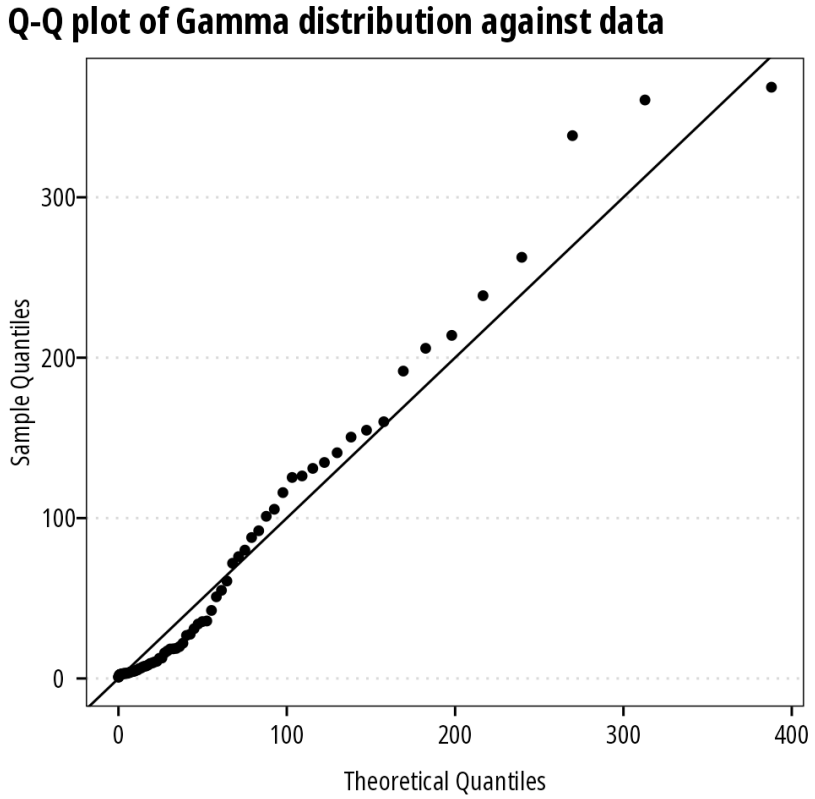


Figure 4: Boxplots of dissolved oxygen concentrations at 5 sites.

Probability plots

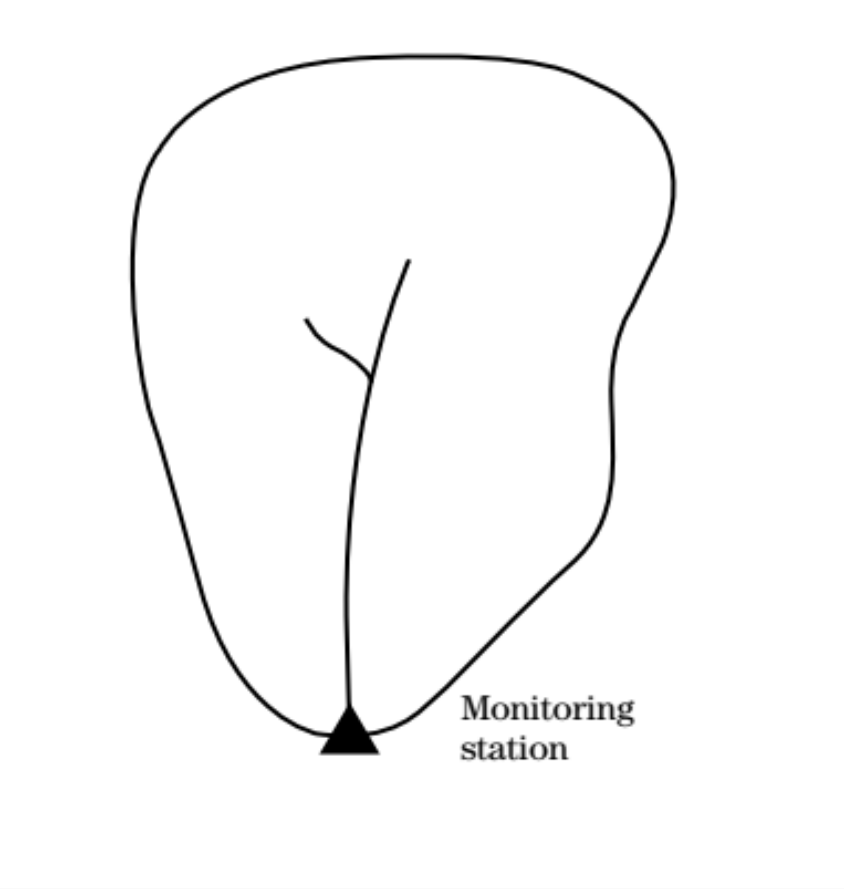
Also called a quantile-quantile (Q-Q) plot. This is the quantile plot generated earlier plotted against quantiles from a theoretical distribution. These are used to evaluate how well the data fits against distributions such as the normal, log-normal, or gamma distribution.





Statistical design for watershed studies

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.

Flow-Weighted Average

Example:

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

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

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

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**.

Permutation Test

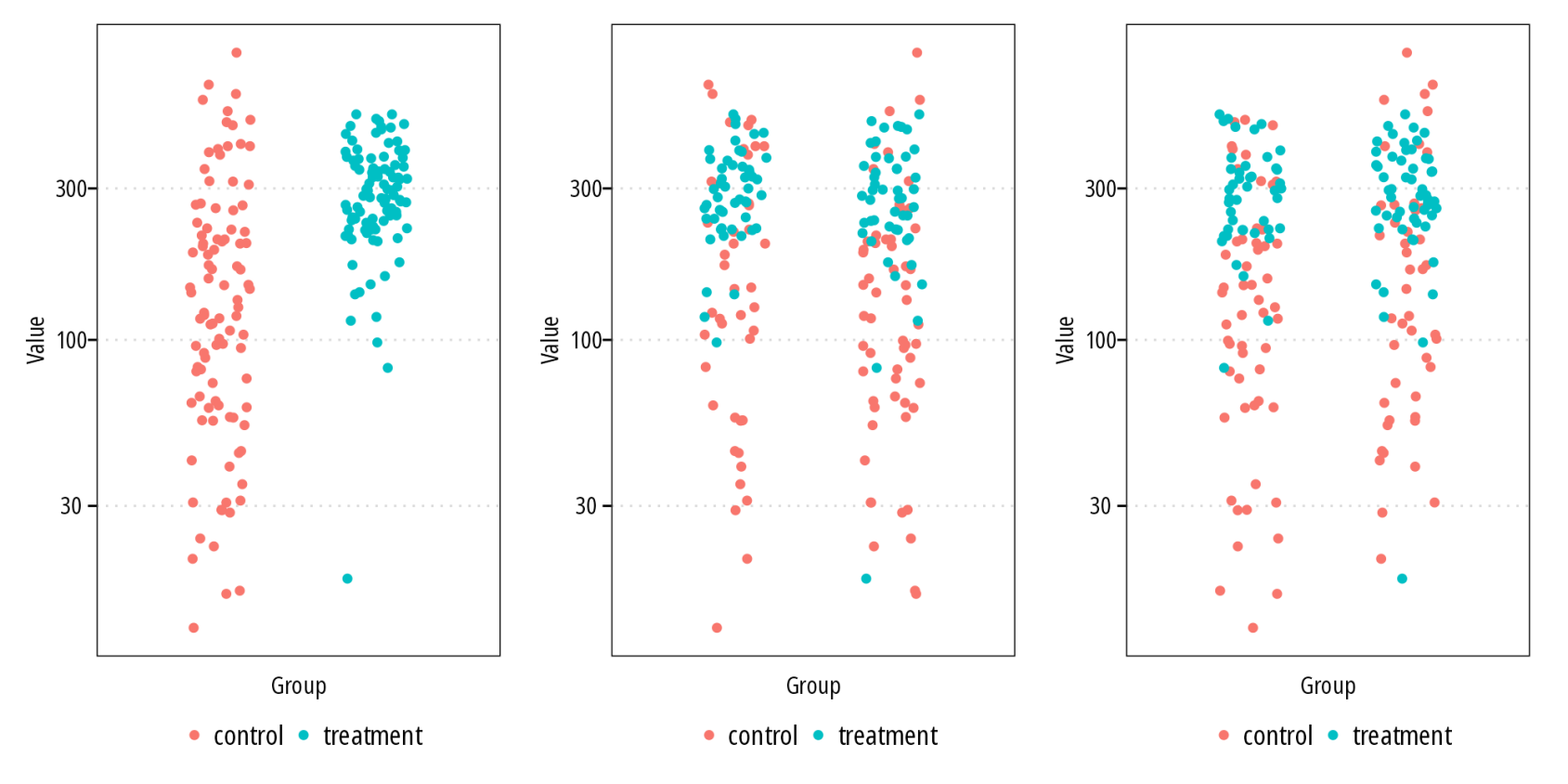
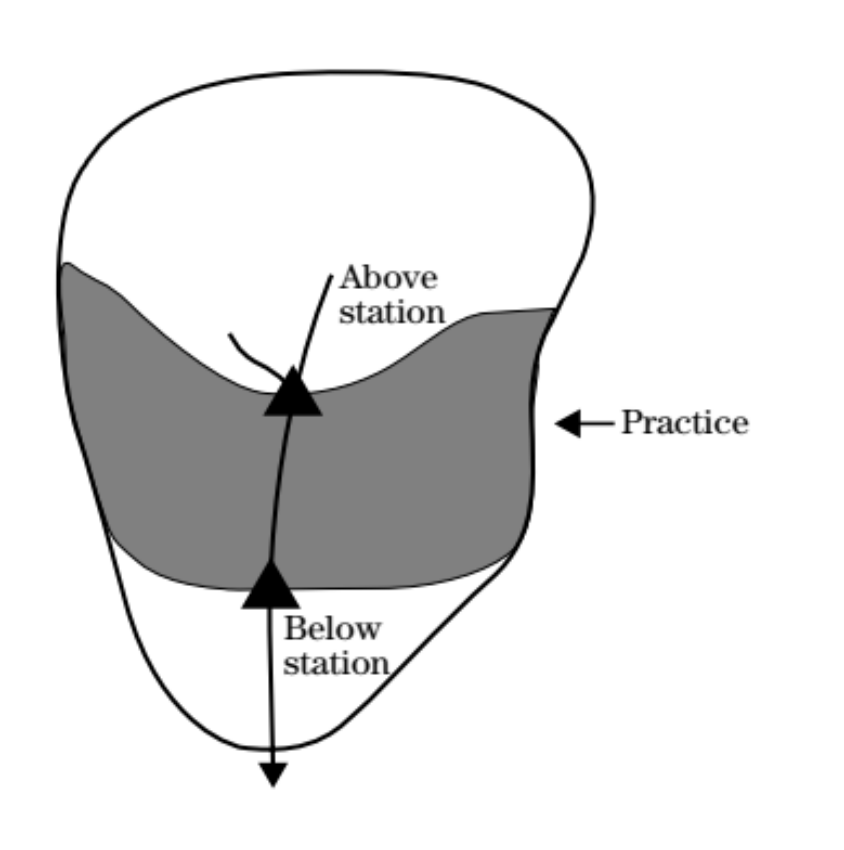


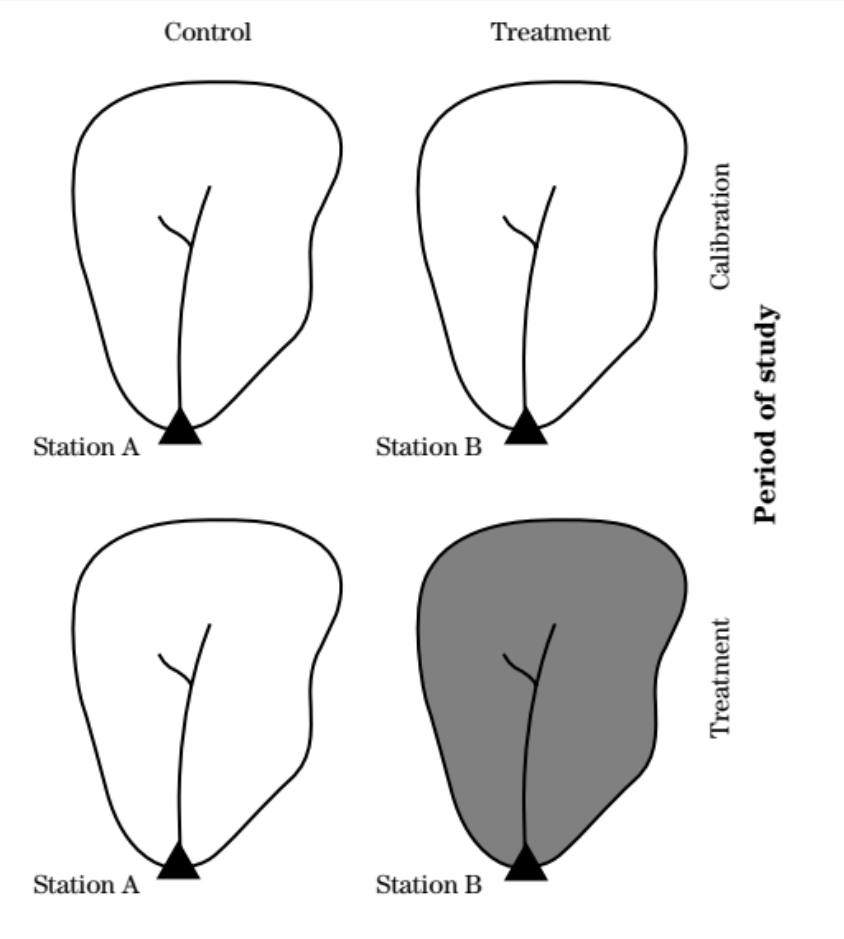
Figure 6: Permutations randomly shuffle that data between groups. Assuming there is no difference between the groups, new reshuffles will have approximately the same differences.

Single watershed study



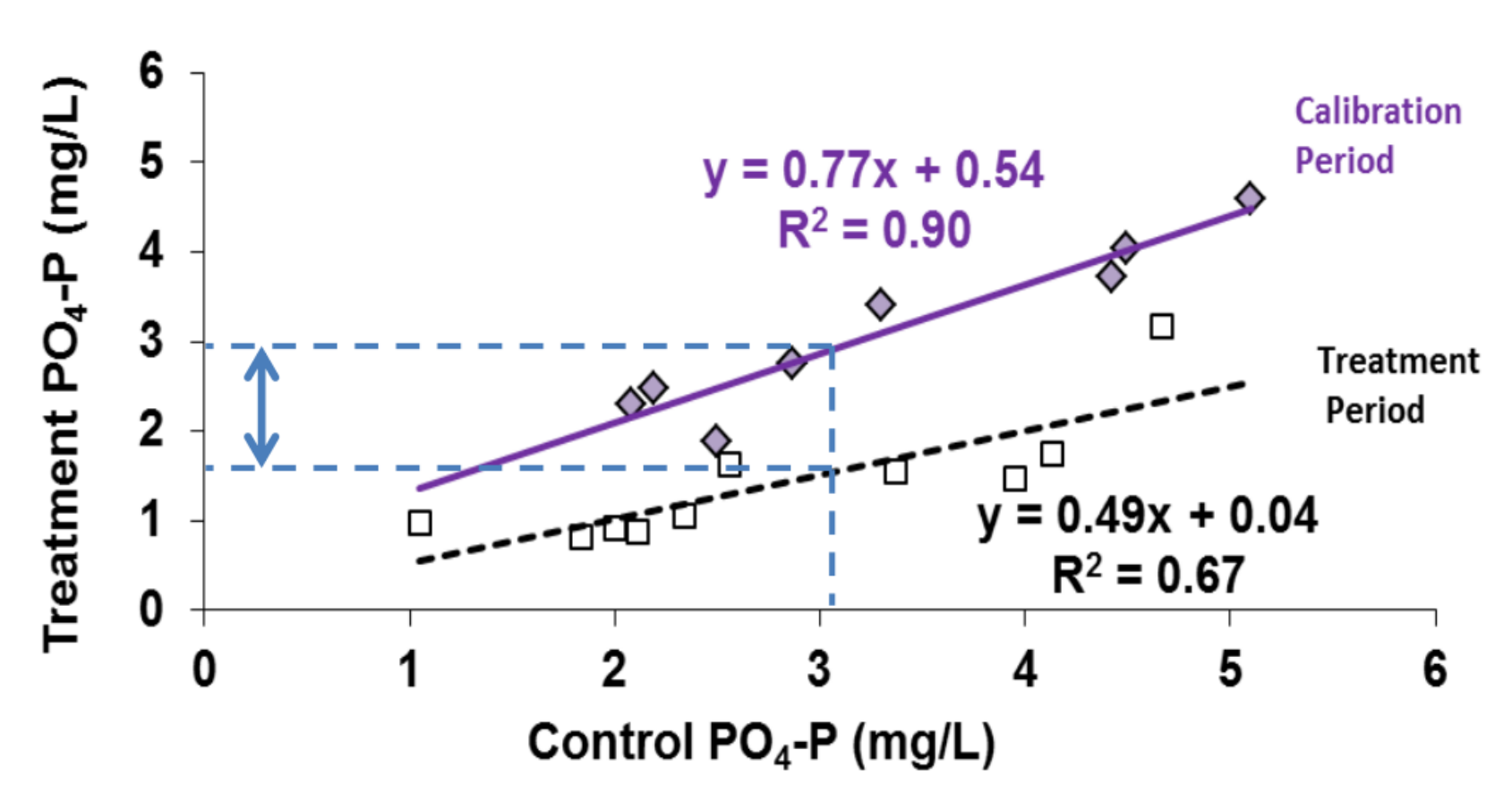
- Compare upstream and downstream concentrations.
- **Parametric paired t-test** or **non parametric signed rank test**.
- Include before and after implementation as factors.
- **Parametric Two factor ANOVA** or **nonparametric Brunner-Dette-Munk test**.
- Permutation tests available for each of these.

Paired Watershed Study



- Two-factor ANOVA, nonparametric Brunner-Dette-Munk test
- Can be setup as a linear regression model
- $Treated = \beta_0 + \beta_1(Control) + \epsilon$

Paired Watershed Study



Source: McFarland and Hauck (2004)

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.

Cheet sheet

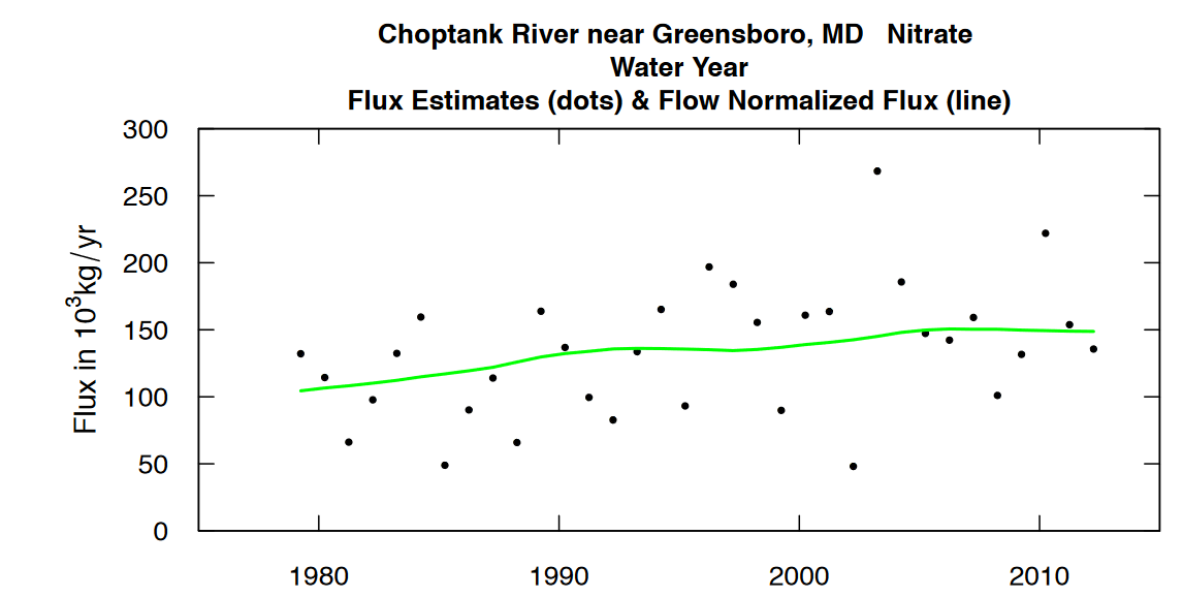
Design	Parametric	Nonparametric	Permutation
Pre/Post	t-test	Rank-sum test	Two-sample permutation test
Upstream/Down-stream	Paired t-test	Signed-rank test	Paired permuatation test
Paired-watershed (BACI)	Two-factor ANOVA or Linear regression	Brunner-Dette-Munk test or generalized linear model	Two-factor permutation test

Trend Analysis

Type	Unadjusted for variable X	Adjusted for variable X
Parametric	Regression of Y on T	Regression of Y on X and T
Nonparametric	Mann-Kendall trend test	Mann-Kendall test on residuals from loess regression of Y on X

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

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 Hirsch and De Cicco (2015).

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.

Major references

Helsel, D. R., Hirsch, R. M., Ryberg, K. R., Archfield, S. A., & Gilroy, E. J. (2020). Statistical methods in water resources: U.S. Geological Survey techniques and methods, book 4, chapter A3. USGS. https://doi.org/10.3133/tm4a3

USDA NRCS. (2003). National Water Quality Handbook Part 614. USDA NRCS. https://archive.epa.gov/water/archive/web/pdf/stelprdb1044775.pdf

Hirsch, R. M., & De Cicco, L. A. (2015). **User Guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval- R Packages for Hydrological Data.** In *Book4, chapter A10* (p. 93). http://dx.doi.org/10.3133/tm4A10

Kuhnert, P. M., Henderson, B. L., Lewis, S. E., Bainbridge, Z. T., Wilkinson, S. N., & Brodie, J. E. (2012).**Quantifying total suspended sediment export from the Burdekin River catchment using the loads regression estimator tool.** *Water Resources Research*, *48*(4), W04533. https://doi.org/10.1029/2011WR011080

McFarland, A. M., & Hauck, L. M. (2004).**Controlling phosphorus in runoff from long term dairy waste application fields.** *JAWRA Journal of the American Water Resources Association*, *40*(5), 1293–1304. https://doi.org/10.1111/j.1752-1688.2004.tb01587.x