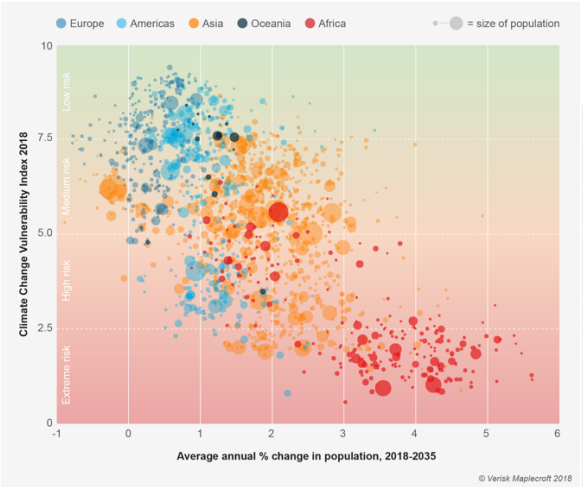
Graphical Data Analysis 01

EN5423 | Spring 2024

w03_graphical_01.pdf
(Week 3)

Contents



Introduction to Graphical Data Analysis

• Graphs are pivotal in statistical analysis, offering insights that might not be evident through statistical measures alone.

• They can reveal patterns, trends, and associations in data, highlighting the potential for misunderstanding if analysis relies solely on numerical summaries.

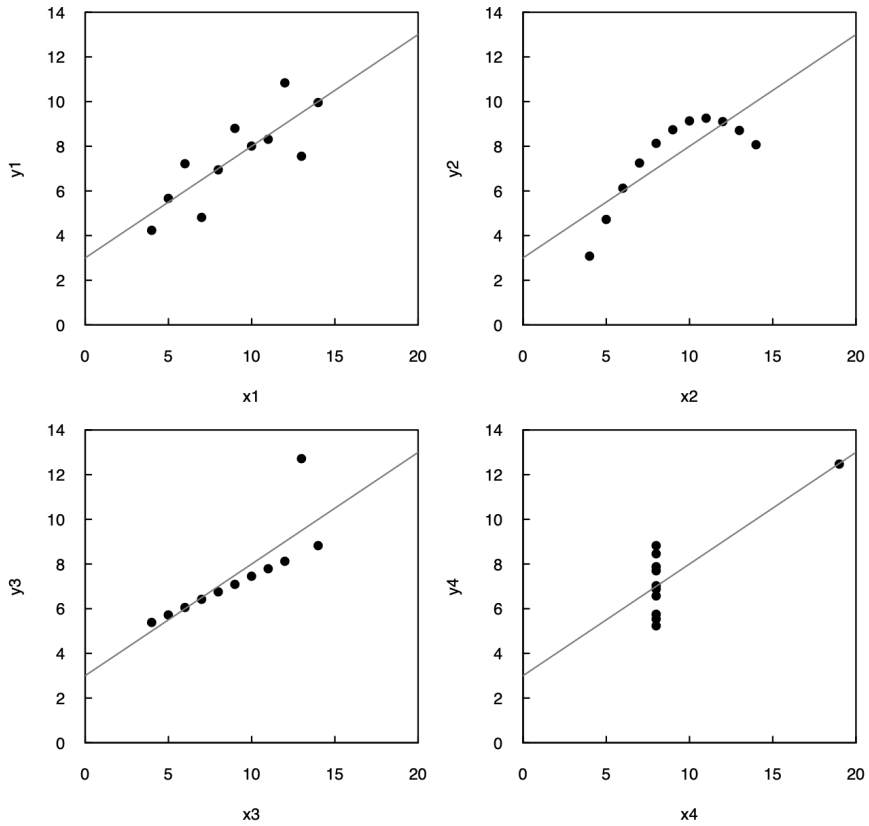


Figure 1. Four scatterplots of datasets that all have the same traditional statistical properties of mean, variance, correlation, and x-y regression intercept and coefficient. These datasets are known as Anscombe's quartet (Anscombe, 1973).

• **Key Roles of Graphs**

1. **Exploratory Data Analysis (EDA):** Graphs aid in understanding and exploring data, helping to identify patterns and generate hypotheses about how systems behave.
2. **Presentation of Results:** Effective communication of analysis findings to others is greatly enhanced through graphical representation.

• **Necessity of Data Plotting**

- The advancement of statistical software has eliminated time constraints as an excuse for not plotting data. Plotting is now an essential step in the analytical process.

- **Graphical Methods Covered**
  - This lecture introduces graphical methods for analyzing single datasets, with a focus on boxplots and probability plots.
  - It discusses comparison methods for multiple groups of data and introduces bivariate plots, including scatterplots with smoothes for enhancing visual interpretation.
  - It concludes with approaches for analyzing multivariate data.

- **Datasets for Demonstration**
  - Two specific datasets are used to illustrate the effectiveness of graphical methods: annual streamflow in the James River and unit well yields in valleys of Virginia. These examples demonstrate how different graphical methods can be applied to real-world data.

1 Graphical Analysis of Single Datasets

Measures of location, or central tendency, provide essential insights into the typical conditions within an environmental dataset. These measures establish baselines for assessing changes, impacts, and the quality of resources.

1.1 Histograms (*Exercise 1*)

Histograms serve as fundamental tools in graphical data analysis, offering insights into a dataset's central tendency, variability, and symmetry.

• Construction and Purpose:

1) Histograms sort data into categories of equal width, displaying the frequency of observations within each category.
2) They approximate the probability density function of the population as sample size approaches infinity.
3) The height of each bar represents either the count (n_i) or the proportion ($\frac{n_i}{n}$) of observations in each category.

• Determining the Number of Bins:

1) The appropriate number of intervals (k) is crucial for effective histogram representation. A recommended approach is to choose k so that 2k≥n2k≥n, where n is the sample size.

• Influence of Bin Selection::

1) The visual interpretation of histograms significantly depends on the chosen number of categories (bins).
2) Different bin widths and centers can lead to varying impressions of data distribution, as illustrated by comparison of figures with different bin settings.

• Usage and Limitations:

1) Histograms provide a general impression of data distribution but are not suitable for precise analysis of individual values.
2) The method of categorizing continuous data into discrete bins can sometimes obscure distribution characteristics.
3) They are particularly effective for presenting discrete data and making distributions understandable to non-specialist audiences.

• Practical Considerations:

1) For preliminary exploration, the default settings of histogram functions (e.g., `matplotlib.pyplot.hist` in Python) usually suffice to generate informative visuals.The method of categorizing continuous data into discrete bins can sometimes obscure distribution characteristics.
2) Customization of bin locations and widths is generally reserved for preparing histograms for presentation or publication.

• Real-world Application:

1) Figure below demonstrates how histograms with different bin configurations can convey varying impressions of the same dataset.

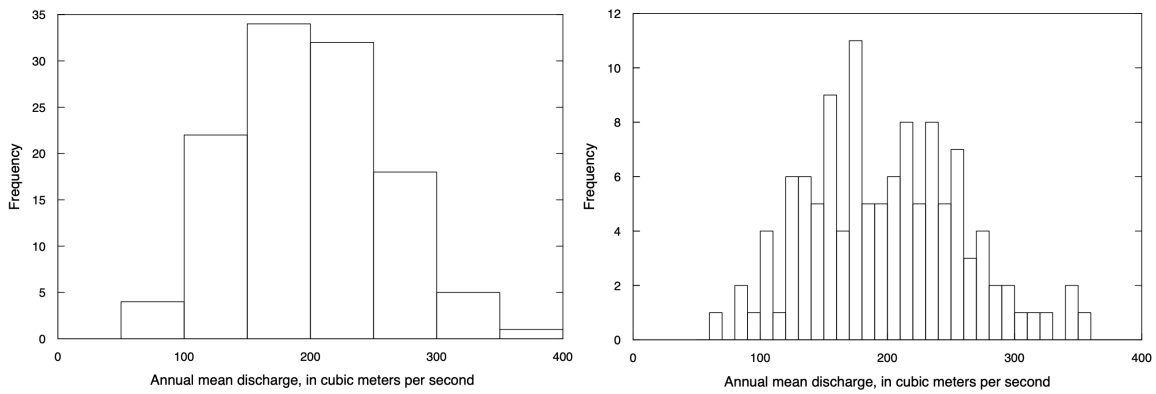


Figure 2. (left) Histogram of annual mean discharge for the James River at Cartersville, Virginia, 1900–2015. Histogram of annual mean discharge for the James River at Cartersville, Virginia, 1900–2015. (right) Annual streamflow data are the same as shown in figure 2.2, but with different interval divisions.

1.2 Quantile Plots (*Exercise 2*) (*HW03 #1*)

• As discussed in the previous section, histograms are a sample-based approximation of a probability density function (pdf).

• Another way to display information about a distribution is to use the integral of the probability density function, which is called the cumulative distribution function (cdf). The cdf is a plot of the probability that the random variable will be less than some specific quantity.

• The vertical scale of a cdf ranges from 0 (for the smallest possible value of the random variable) to 1 (for the largest possible value).

• Quantile plots, also known as empirical cumulative distribution functions (ecdf), offer a graphical representation of the distribution of sample data, based on cumulative probabilities.

• Foundation and Definition:

Quantile plots visualize the cdf, showing the probability that a random variable is less than a specific value. The vertical scale of a cdf plot ranges from 0 to 1, representing the smallest and largest possible values of the random variable, respectively.

• Advantages of Quantile Plots:

1) **No Arbitrary Categories**: Unlike histograms, quantile plots do not require data to be sorted into arbitrary bins.
2) **Displays All Data**: Each data point is shown, offering a more comprehensive view than boxplots.
3) **Distinct Data Points**: Every observation in a quantile plot has a unique position, eliminating overlap.

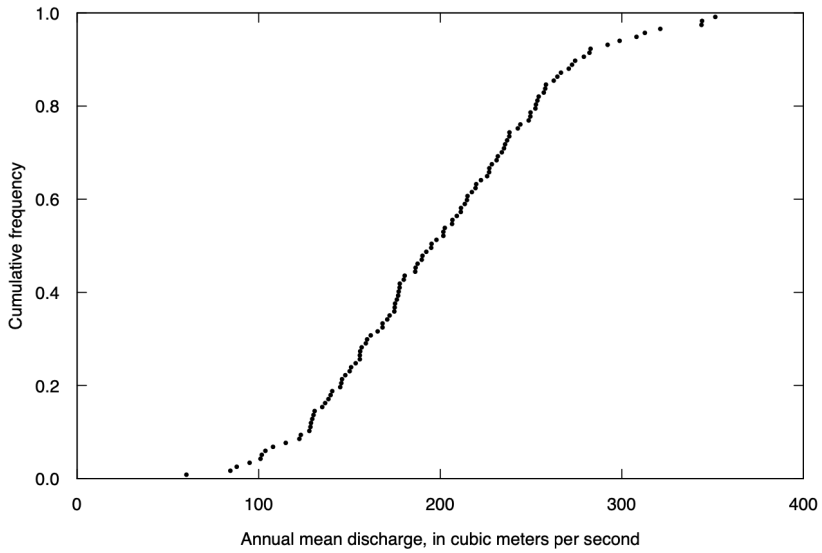


Figure 3. Quantile plot of annual mean discharge data from the James River, Virginia, 1900–2015.

• Construction of Quantile Plots:

1) Data are ranked from smallest to largest, with the smallest assigned a rank of 1 and the largest a rank of n (the sample size).
2) Data values are plotted on one axis (typically the horizontal axis), with the plotting position on the other axis, based on the rank and sample size.
3) The Weibull plotting position formula, $p_i = \frac{i}{n+1}$, is commonly used, placing the maximum value's plotting position below 1.0, acknowledging a nonzero probability of exceeding the maximum observed value.

Table 1. Quantile plot values for streamflow data from the James River, Virginia, 1900–2015

[x, annual mean discharge in cubic meters per second ($\frac{m^3}{s}$); p, cumulative frequency values; i, rank of the obervation; dots indicate data not shown in table but included in figure 2.4]

i	x_i	p_i
1	60.2	0.0085
2	84.4	0.0171
3	87.9	0.0256
4	95.0	0.0342
5	101.0	0.0427
.	.	.
.	.	.
112	312.7	0.9573
113	321.2	0.9658
114	343.9	0.9744
115	344.2	0.9829
116	351.5	0.9915

Example 1

• Make a quantile plot for the following squential data:

[1, 4, 2, 3, 4, 5, 6, 8, 10, 12, 13, 15, 20, 30, 13, 15, 16, 3, 9, 10]

>>

• Real-world Application and Example:

An example quantile plot of streamflow data illustrates how quantile plots can reveal distribution details such as skewness, spread, and extreme values with greater precision than histograms.

• Variations of quantile plots are used for three purposes:

1) To compare two or more data distributions (a Q-Q plot).
2) To compare data to a normal distribution (a normal probability plot, a specialized form of the Q-Q plot).
3) To calculate frequencies of exceedance (for example, a flow-duration curve used to evaluate streamflow data).

• Choice of Plotting Position Formulas:

1) Various formulas exist for calculating plotting positions, including Weibull (used in hydrology for flow-duration and flood-frequency curves) and Blom (used for normal probability plots).
2) The choice of formula can depend on tradition, the specific needs of the analysis, or the defaults of statistical software packages.
3) Historically, in the field of hydrology and in statistics in general, different plotting positions have been used to construct quantile plots. The choice of plotting positions used depends on these purposes, but also on tradition or on automatic selection of methods in various statistical software packages. Most plotting positions have the general formula $p = \frac{i-\alpha}{n-\alpha-\beta+1}$ where α and β are constants. Commonly used formulas are listed in table 2.

• Significance in Hydrology and Statistics:

The Weibull formula is highlighted for its realism in acknowledging the probability of exceeding maximum observed values, making it preferred for hydrological analyses and flood frequency determinations in the United States.

Table 2. Definitions and comments on eight possible plotting position formulas, based on Hyndman and Fan (1996) and Stedinger and others (1993).

[NA, not applicable; i, rank of the observation; n, sample size; p_i, calculated probability for the i^{th} ranked observation; p_n, rank of the largest observation]

Reference	α	β	Formula p_i=	Type in R quantile function	Comments
Parzen (1979)	0	1	i/n	4	p_n = 1.0, poor choice: suggests largest observation can never be exceeded
Hazen (1914)	1/2	1/2	$(i-(1/2))/n$	5	Traditional in hydrology
Weibull (1939), also Gumbel (1958)	0	0	$(i)/(n+1)$	6	Unbiased exceedance probabilities
Gumbel (1958)	1	1	$(i-1)/(n-1)$	7	p_n = 1.0, poor choice: suggests largest observation can never be exceeded
Reiss (1989)	1/3	1/3	$(i-(1/3))/(n+(1/3))$	8	Median unbiased quantiles
Blom (1958)	3/8	3/8	$(i-(3/8))/(n+(1/4))$	9	Unbiased quantiles for normal
Cunnane (1978)	2/5	2/5	$(i-(2/5))/(n+(1/5))$	NA	Approximate quantile unbiased
Gringorten (1963)	0.44	0.44	$(i-0.44/(n+0.12))$	NA	Optimized for Gumbel distribution

1.3 Boxplots (*Exercise 3*) (*HW03 #2*)

• Boxplots are succinct graphical displays that summarize the distribution of a dataset effectively.

• Key Features of Boxplots:

1) **Center of Data:** Represented by the median, marked as a centerline within the box.
2) **Variation or Spread:** Shown through the interquartile range (IQR), the height of the box.
3) **Skewness:** Indicated by the quartile skew, observed as the relative size of the box halves.
4) **Outliers:** Identified as individual symbols beyond the whiskers, indicating unusual values.

• Advantages for Data Comparison:

1) Boxplots are particularly useful for comparing these statistical attributes across multiple datasets.

• Elements of a Boxplot:

1) **Box:** Represents the middle 50% of the data, between the 25th and 75th percentiles.
2) **Hinges:** The top and bottom of the box, approximating the 75th and 25th percentiles, respectively.
3) **Median Line:** A line within the box indicating the sample median.
4) **Whiskers:** Extend from the box to show the range of the data, defined precisely by a maximum length from the box (typically no more than 1.5 times the IQR).
5) **Outside Values:** Data points lying beyond the whiskers, marked individually, indicating outliers.

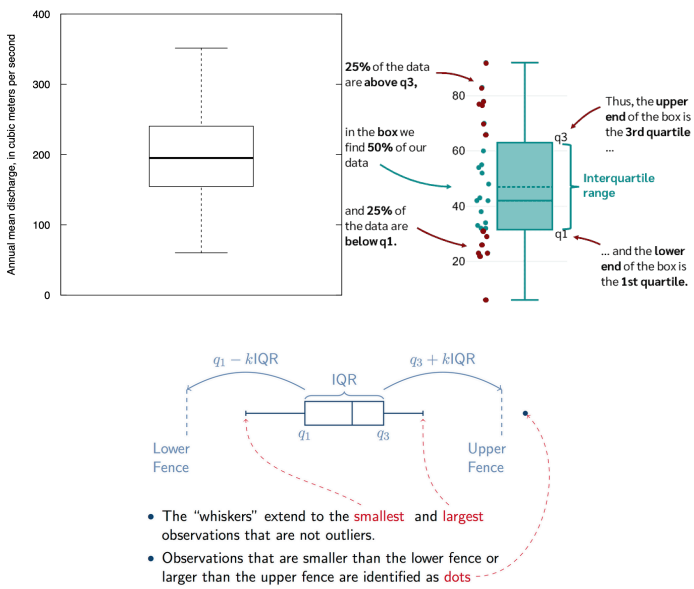


Figure 4. (upper left) A boxplot of annual mean discharge values from the James River, Virginia, 1900–2015

• **Whisker and Outlier Definitions:**

By default, whiskers extend to the most extreme data points within 1.5 times the box's length. Data beyond this are considered outside values or outliers. This choice is related to the expected distribution of observations in a large normal sample, where approximately 5% are expected to be outliers.

• **Interpretation of Boxplots:**

While boxplots provide a comprehensive summary of data distribution, including symmetry and outliers, they may not reveal characteristics like bimodality. The distribution within the interquartile range might remain obscured within the box.

• **Real-world Application:**

Examples given include annual mean discharge data for the James River, demonstrating the boxplot's ability to indicate symmetry and outlier presence or absence.

1.4 Probability (Q-Q) Plots (*Exercise 4*)

Q-Q (Quantile-Quantile) plots are graphical tools used to compare two sets of quantiles against each other. Typically, one set represents the quantiles from a sample dataset, plotted on one axis, while the other set may represent either a theoretical distribution's cumulative distribution function (cdf) or the quantiles from another empirical distribution. ***These plots are particularly useful for assessing how well a sample fits a theoretical distribution***, such as normal, lognormal, or gamma distributions.

• **Types of Q-Q Plots:**

1) **Probability Plot:** A specific type of Q-Q plot where the sample data's quantiles are compared against the quantiles of a theoretical distribution. This plot aims to assess the fit between the sample data and the theoretical model.
2) **Empirical Distribution Comparison:** Another Q-Q plot type involves comparing the quantiles of two empirical distributions to evaluate their similarity.

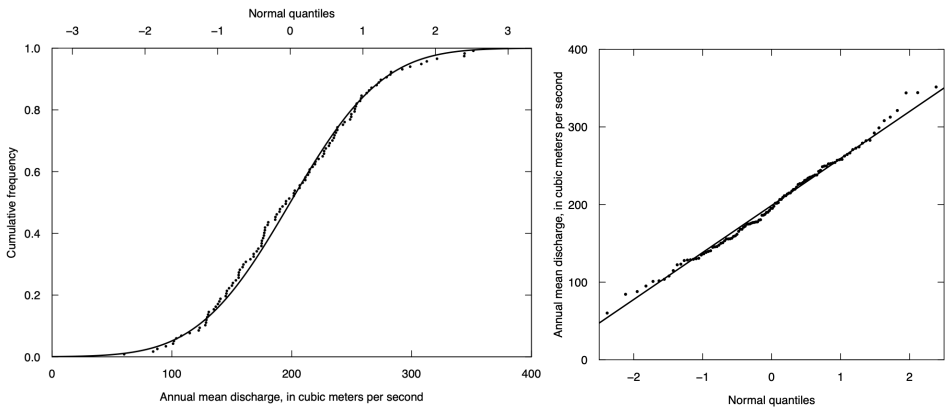


Figure 5. (left) Overlay of James River annual mean discharge (1900–2015) and standard normal distribution quantile plots. (right) Probability plot of James River annual mean discharge data (1900–2015).

• Purpose and Application:

1) **Distribution Fit Assessment:** Probability plots are instrumental in determining how closely sample data align with expected theoretical distributions. This assessment can be somewhat crudely performed by comparing histograms of sample data with the probability density function of theoretical distributions. However, Q-Q plots offer a more visually intuitive means by evaluating the linearity of data points against a theoretical straight line.
2) **Advantages Over Histograms:** Research suggests that human perception more readily identifies deviations from straight lines than from curves. Therefore, Q-Q plots, by transforming theoretical distributions into straight lines, make deviations more apparent.

• Construction and Interpretation:

1) **Normal Probability Plot Example:** An example provided is the normal probability plot of the James River streamflows. This plot demonstrated a close adherence to a straight line, suggesting a strong consistency with a normal distribution, with minor deviations in the upper tail indicating slight extremity.
2) **Practical Construction:** To construct a probability plot, sample quantiles are re-expressed as standard normal quantiles. A straight line is defined by the sample mean and standard deviation. Individual data points are plotted such that a normal distribution will present as a straight line (**Figure 5** (right))

• Methodology:

The construction involves plotting nn observations sorted from smallest to largest against the inverse of the CDF for the standard normal distribution (mean = 0, SD = 1). The methodology can be summarized with the following equations:

1) The line in a normal Q-Q plot is defined as:

$$Q = \bar{Q} + z \cdot s_Q \tag{1}$$

where $\bar{Q}$ the sample mean of the Q_i values, s_Q is the sample standard deviation of the Q_i values, and z is the z-score from the ***standard normal distribution*** corresponding to the cumulative probability.

2) The plotting positions for the *i*-th data point are calculated as:

$$p_i = \frac{i}{n+1} \tag{2}$$

where *n* is the total number of observations.

3) The z-score (Z_i) for the ii-th data point in the plot is determined by::

$$Z_i = F^{-1}(p_i) \tag{3}$$

where F^{-1} is the inverse of the cumulative distribution function for the standard normal distribution, essentially the `scipyt.stats.norm.ppf` function in Python.

• Significance:

Q-Q plots provide a nuanced, visually intuitive method for assessing sample data's adherence to theoretical distributions. Their utility extends beyond mere visual comparison offered by histograms, offering a clearer depiction of how sample data conforms to expected distribution patterns, thereby facilitating a deeper understanding of data characteristics relative to theoretical models.

1.5 Q-Q plots as Exceedance Probability Plots

In the context of water resources management, it is common to enhance a probability plot with an additional horizontal axis, which represents the probability of exceedance. This adaptation essentially reinterprets the quantile information, allowing for the omission of the standard quantile scale. Such modifications result in a plot that, while fundamentally identical to a conventional probability plot, includes an added layer of insight through the exceedance probability scale. This adjustment facilitates a more intuitive understanding of the graphical data. An illustration of this concept could be seen in a hypothetical **Figure 6**, mirroring the structure of a previously discussed probability plot (e.g., **Figure 5**) but with the critical enhancement of an exceedance probability axis. This methodology, versatile across various distributions, was historically executed using specially designed probability paper before the advent of statistical software, enabling normal distributions to be depicted as straight lines, thereby simplifying the visualization and interpretation process for hydrologists.

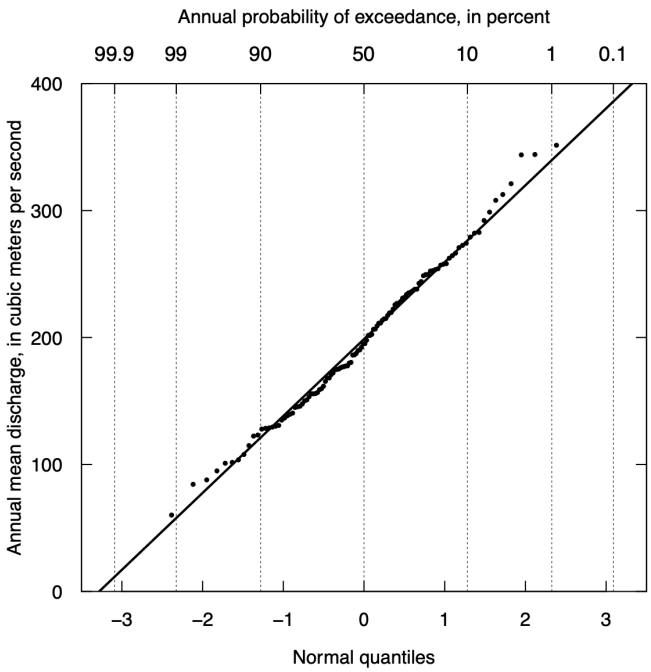


Figure 6. Exceedance probability plot for the James River annual mean discharge data (1900–2015).

1.6 Deviations from a Linear Pattern on a Probability Plot

• Deviations from a linear pattern in a probability plot are indicative of how a dataset's distribution diverges from a theoretical distribution. Such deviations are crucial for identifying the nature and extent of these departures, offering insights into:

1) **Overall Asymmetry or Skewness:** How the distribution leans towards one side, affecting the tail's stretch further from the mean in one direction compared to the other.
2) **Presence of Outliers:** Extreme values that fall far from the bulk of the data distribution.
3) **Kurtosis (Heaviness of the Tails):** Whether the tails of the distribution are heavier (more outliers than expected) or lighter (fewer outliers) compared to a normal distribution.

• Contrastingly, boxplots of the same dataset may not clearly show this asymmetry due to their design focusing on quartile ranges rather than detailed distribution shapes. However, boxplots can effectively indicate pronounced skewness through the visual disparity of their whiskers or the positioning of their median.

• In essence, probability plots serve as a visual diagnostic tool for assessing the distributional characteristics of environmental data, guiding the selection of appropriate analytical or transformation techniques to better understand and model these data.

1.7 Probability Plots for Comparing Among Distributions

• Probability plots can be generated for ***any theoretical distribution***, not just the normal distribution. These plots provide a visual method to assess the fit of data to a theoretical distribution by comparing data quantiles to a theoretically linear relationship.

• The closer the data aligns with this straight line, the more appropriate the theoretical distribution is considered for the data. This concept is extended into a formal hypothesis test, the probability plot correlation coefficient test, which is detailed in later discussions.

• While hydrology extensively explores distribution selection and parameter estimation for analyzing flood and low flows, this text briefly touches on these topics. For more comprehensive insights into frequency analysis using probability plots, works by Vogel (1986), Vogel and Kroll (1989), and Stedinger et al. (1993) are recommended.

2 Graphical Comparisons of Two or More Datasets

This section delves into the use of graphical methods for ***comparing multiple datasets***, highlighting the effectiveness and limitations of different types of plots. While histograms, boxplots, and probability plots (including quantile-quantile or Q-Q plots) have been discussed for ***individual dataset analysis***, their utility varies significantly when applied to comparisons between datasets.

- **Histograms**

While capable of displaying distributional information for single datasets, histograms fall short in providing clear, detailed comparisons across multiple datasets. The overlapping nature of histograms can obscure differences and similarities between groups, especially when dealing with multiple overlays or dense datasets.

- **Boxplots**

Boxplots are highly effective for comparing distributional characteristics across several groups of data. They excel in clarity and the ability to easily distinguish between key features such as medians, quartiles, and the presence of outliers. This makes boxplots an ideal choice for visualizing and comparing the distributions of multiple datasets simultaneously.

- **Quantile-Quantile (Q-Q) Plots**

Q-Q plots offer a nuanced comparison between two datasets by mapping their quantiles against each other. This method is particularly informative for assessing how similarly two datasets are distributed, providing insights into their relative skewness, kurtosis, and overall distributional shape compared to one another.

The practical application of these graphical methods is illustrated through the comparison of unit well yields from two distinct geological settings in Virginia, as studied by Wright (1985). The datasets comprise measurements ***from 13 wells in valleys underlain by fractured rocks*** and ***12 wells in valleys with unfractured rocks***. These comparisons serve to demonstrate the strengths and weaknesses of each graphical method in highlighting differences between small datasets characterized by underlying geological variations.

Keypoints:

1) While histograms might provide basic distributional insights, boxplots and Q-Q plots offer more precise and informative comparisons between groups of data.
2) The choice of graphical method depends on the specific aspects of the datasets one aims to compare, with boxplots generally offering the best clarity for multiple group comparisons and Q-Q plots providing deeper insights into the distributional relationship between pairs of datasets.

2.1 Histograms

- These histograms showcase the right-skewness of each dataset.

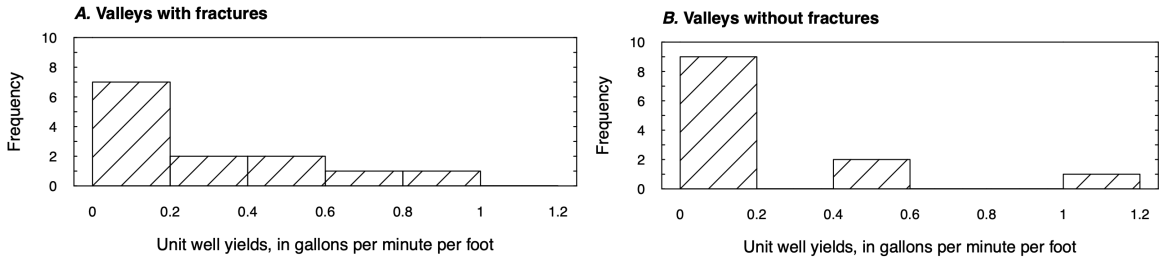


Figure 7. Histograms of unit well yield data for (A) valleys with fractures, and (B) valleys without fractures.

- Histograms reveal the right-skewness of well yield data from different geological settings but struggle to clearly differentiate between the two datasets.

• Despite uniform bin sizes and sequential plotting to enhance comparability, histograms fall short in effectively comparing distribution centers and spreads.

• Superimposing histograms on a single axis, intended for direct comparison, often results in confusing and uninformative visuals, making it a less preferred method.

2.2 Dot-and-line Plots of Means and Standard Deviations

• Dot-and-line plots, while used for dataset comparison, are notably less informative than boxplots, primarily highlighting mean yields without adequately depicting other distributional characteristics.

• These plots assume symmetry, leading to unrealistic representations (e.g., negative values for physically impossible scenarios) and fail to provide insights into data symmetry or the presence of outliers.

• Limited information on data spread is offered, with the standard deviation bars not necessarily reflecting the true distribution range, especially in skewed datasets with outliers; side-by-side boxplots are recommended for a more comprehensive comparison.

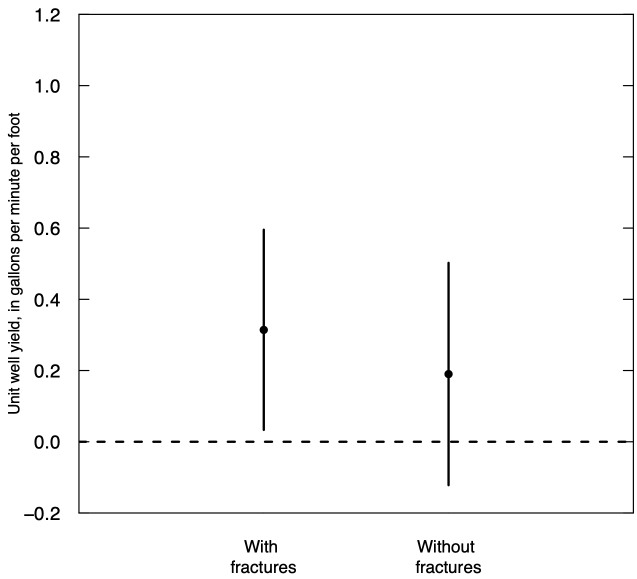


Figure 8. Dot-and-line plot of the unit well yield datasets for areas underlain by either fractured or unfractured rock.

2.3 Side-by-side Boxplots (*Exercise 5*) (*HW03 #3*)

• Side-by-side boxplots effectively highlight distributional differences in well yield data, showing higher median yields and slightly larger interquartile ranges (IQRs) for wells with fractures compared to those without, while indicating similar maximum values and right-skewness in both datasets.

• The mean yield, especially for wells without fractures, appears inflated due to skewness, suggesting that actual differences between groups may be more significant than mean values

alone indicate; boxplots' ability to reveal central tendency, variability, and skewness makes them superior for comparing datasets.

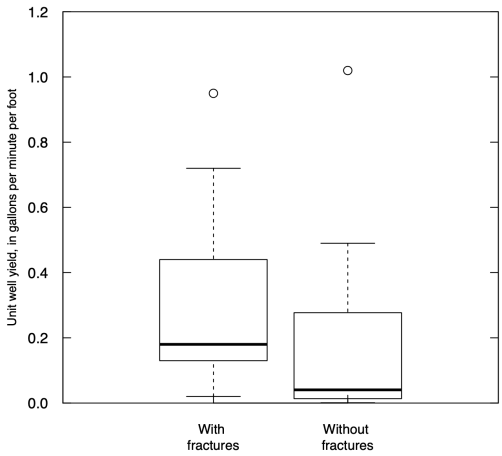


Figure 9. Side-by-side boxplots of the unit well yield datasets for areas underlain by either fractured or unfractured rock.

• Such boxplots concisely display key data characteristics across multiple groups, as demonstrated in an example comparing ammonia nitrogen concentrations in different sections of the Detroit River, effectively summarizing changes in water quality in a compact visual format.

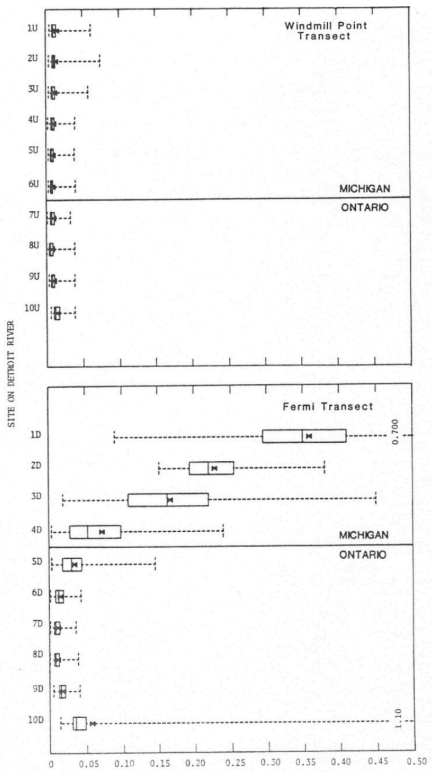


Figure 10. Boxplots of ammonia nitrogen concentrations as a function of location on two transects of the Detroit River (from Holtschlag, 1987). The Windmill transect lies upstream of Detroit and the Fermi transect lies downstream of Detroit.

• Side-by-side boxplots effectively illustrate seasonal variations in dissolved nitrate-plus-nitrite concentrations in the Illinois River, highlighting significant differences across months with higher concentrations observed from December through June and lower levels during summer and fall.

• The boxplots indicate that winter and spring months show relatively high nitrate concentrations with little variance among them, while concentrations drop in the summer, reaching the lowest points in August and September. ***These seasonal patterns reflect the influence of nitrogen fertilizer application timings and biological processes on nitrate levels***.

• This example underscores the complex seasonal patterns of hydrologic variables like nitrate concentrations, which are influenced by a combination of physical processes (e.g., rainfall, snowfall), biological activity, and human agricultural practices, rather than following simple sinusoidal temperature patterns.

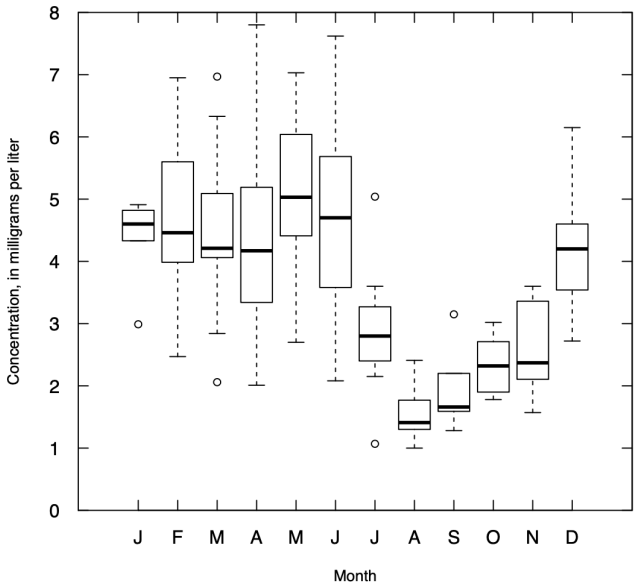


Figure 11. Side-by-side boxplots by month for dissolved nitrate plus nitrite for the Illinois River at Valley City, Illinois, water years 2000–15.

2.4 Q-Q Plots of Multiple Groups of Data

• Side-by-side boxplots effectively highlight distributional differences in well yield data, showing higher median yields and slightly larger interquartile ranges (IQRs) for wells with fractures compared to those without, while indicating similar maximum values and right-skewness in both datasets.

• Q-Q plots can compare two empirical distributions, allowing for a detailed assessment of each quantile, not just quartiles as in boxplots.

• In the comparison of well yield data, Q-Q plots revealed right-skewness and showed that wells with fractures generally have higher yields than those without, particularly for the lower 75% of data.

• Q-Q plots can identify if datasets differ by additive or multiplicative constants, providing insights into underlying relationships. A Q-Q plot comparing well yields from fractured versus unfractured areas (**Figure 12**) shows that the lower 75% of data points have yields in fractured areas roughly 4.4 times higher than those in unfractured areas, as indicated by a slope greater than 1 and not parallel to the reference line $Y_p = X_p$.

• For the highest quantiles, the yields between the two areas converge towards equality, suggesting that for higher yielding wells, the presence of fractures may have less influence on well yield.

• These patterns might inform hydrological interpretations, such as the effect of well depth on yield or potential data misclassification, highlighting the Q-Q plot as an insightful tool for preliminary data analysis before conducting hypothesis testing.

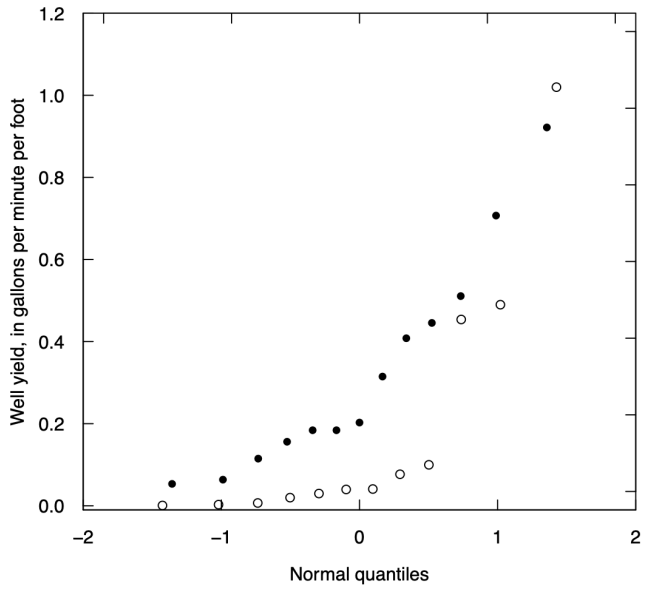


Figure 12. Probability plots of the unit well yield data. Solid circles, wells located in areas of fractured rock; open circles, wells located in areas without fracturing.