Lecture 11
Exploratory Data Analyses

Chao Song

College of Ecology

Lanzhou University

October 27, 2024

What is statistics?

The **population** refer to the entire group of individuals or items that a study is interested in. In mathematical terms, it is the collection of all possible observations of a random variable. A subset of the population is a **sample**.

A **parameter** is a fixed numeric value that describes a population's characteristics. A **statistic** is a number calculated from a sample and is commonly used to estimate or make inference about a parameter.

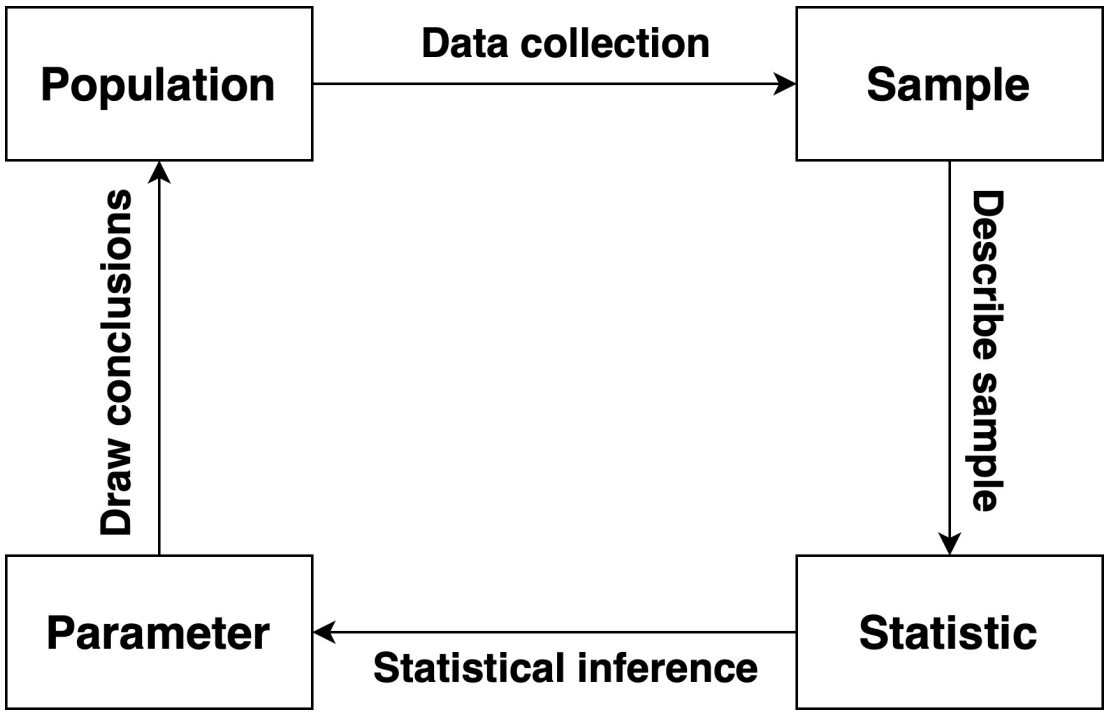
What is statistics?

Example: We are interested in quantifying the average plant richness in alpine grassland in Tibetan Plateau. We thus made a field survey where we randomly selected 20 alpine grassland sites in Tibetan Plateau and measured plant richness. What is the population and what is the sample?

The population is plant richness in all alpine grasslands in Tibetan Plateau. The 20 sites we surveyed are the sample. The mean plant richness across all alpine grasslands in Tibetan Plateau is the parameter we try to quantify. The sample mean richness from the 20 sites is a statistic we used to estimate the true population mean richness.

What is statistics?

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of the data. Two essential tasks of statistics are **parameter estimation** and **hypothesis testing**.



Descriptive statistics

Before making inferences from the data, it's essential to examine the data to

- catch mistakes;
- see patterns in the data;
- find violations of statistical assumptions;
- to generate hypothesis

To examine the data, we typically

- calculate numeric summaries of the data;
- use figures to visualize its distribution.

Numerical summaries of data

We often numerically summarize the data in several aspects, including

- **Central tendency**: they are computed to give a “center” around which the data are distributed;
- **Variability**: they describe the spread of the data, or how far away the data are from the center;
- **Distribution**: they describe the shape of the distribution.

Mean

Arithmetic mean, often referred to as the **sample mean**, is the most commonly used measure of central tendency. Suppose we have a random sample $x_1, x_2, \ldots, x_n$, the sample mean $\overline{x}$ is

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Other types of mean

Geometric mean is defined as

$$\overline{x} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

Example: In a discrete population growth model, changes in population size is modeled as $N_{t+1} = r_t N_t$. Thus, $N_t = (\prod_{i=1}^{t} r_i) N_0$. In this case, the mean population growth rate is best represented by the geometric mean:

$$\overline{r} = (r_1 r_2 \cdots r_t)^{\frac{1}{n}}$$

Other types of mean

Weighted mean is defined as

$$\overline{x} = \sum_{i=1}^{n} w_i x_i,$$

where w_i are weights and $\sum_{i=1}^{n} w_i = 1$.

Example: we measured green house gas emission rate f from n lakes, each with surface area of s_i, we want to calculate the mean green house gas emission rate from these lakes. In this case, the mean weighted by each lake's surface area best represented the mean emission rate.

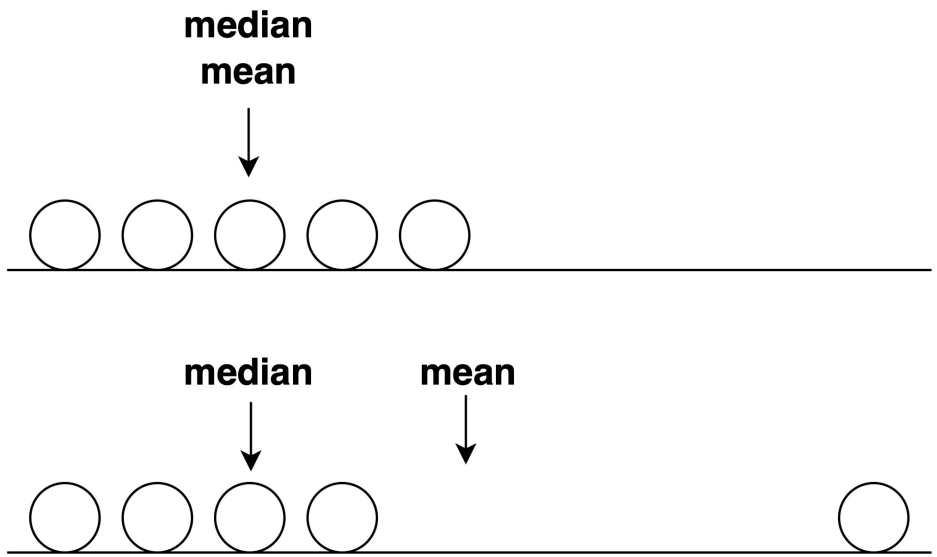
$$\overline{f} = \sum_{i=1}^{n} w_i f_i, \quad w_i = s_i / \sum_{i=1}^{n} s_i$$

Median

For a distribution, the **median** $q_{0.5}$ is defined as $F(q_{0.5}) = 0.5$. For a sample:

- if there are odd number of observations, find the middle value;
- if there are even number of observations, find the middle two and average them.

Median and mean both measures the location of a distribution. An important property of median is that it is less sensitive to outliners or extreme values.



Variance and standard deviation

Sample variance measures the spread of the data and is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Sample standard deviation is the square root of sample variance. Standard deviation has the same unit as the data.

Empirical rule: Let $x_1, x_2, \ldots, x_n$ have a sample mean $\overline{x}$ and a standard deviation s. If the data are from a distribution that is roughly normal, then for large samples:

- approximately 68% of data are within $\overline{x} \pm s$;
- approximately 95% of data are within $\overline{x} \pm 2s$;
- approximately 99.7% of data are within $\overline{x} \pm 3s$;

Percentiles

For a distribution, the *pth* percentile q_p is defined as $F(q_p) = p$ or equivalently $P(x \leqslant q_p) = p$. For a sample, the *pth* **sample percentile** has approximately np observations less than it and $n(1-p)$ observations greater than it.

To achieve this in a sample, we take the $(n+1)p$ smallest value as the sample percentile provided that $(n+1)p$ is an integer. If it is not an integer but is equal to r plus some fraction, we used a linear interpolation between the *rth* and $(r+1)st$ value.

$$q_p = x_r + p(x_{r+1} - x_r)$$

Percentiles

Example: Suppose we have a sample of 100 observations. If $x_1, x_2, \ldots, x_n$ denote the ordered observations from the smallest to the largest, how is the 25% percentile calculated?

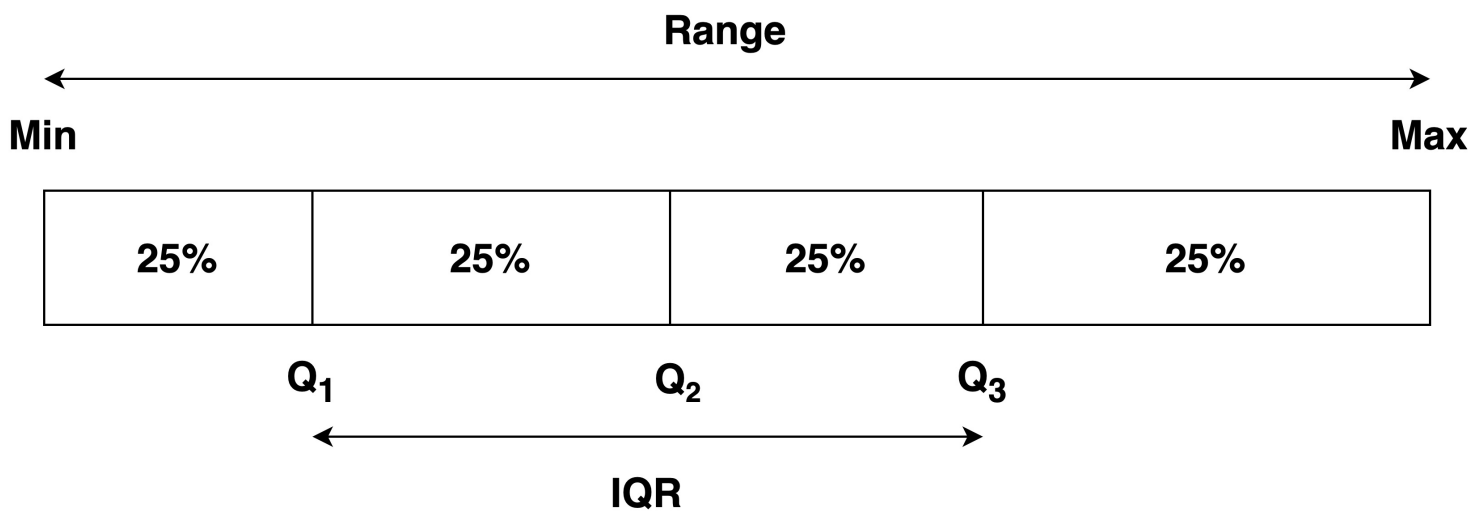
For a sample of size 100, $q_{0.25}$ should have 25 observations less than it and 75 observations greater than it. Here, $(n+1)p = (100+1) \times 0.25 = 25.25$, thus, $q_{0.25}$ is

$$q_{0.25} = x_{25} + 0.25 \times (x_{26} - x_{25})$$

Range, quartile, and IQR

Statistics commonlu used to describe the distribution of the data include

- **Range**: maximum – minimum;
- **Quartiles**: the 25%, 50%, and 75% percentile of the data, quartiles are often denoted Q_1, Q_2, and Q_3;
- **Interquartile range (IQR)**: $Q_3 - Q_1$.



Numeric summaries of data

You cannot fully understand the data from numeric summaries alone. The well-known "**Simpson Paradox**" illustrate this issue.

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

(UC Berkeley gender bias study is a well known example of Simpson's Paradox)

Data visualization

While descriptive statistics summarize certain characteristics of the data, they do not show the whole picture. Visualization of the data is usually necessary.

How you visualize the data depends on the purpose of analysis and more importantly, the type of data. For visualizing univariate data:

- Categorical: bar chart or pie chart
- Ordinal: bar chart or pie chart but ordered;
- Numeric: histogram, boxplot, or violin plot.

Histogram

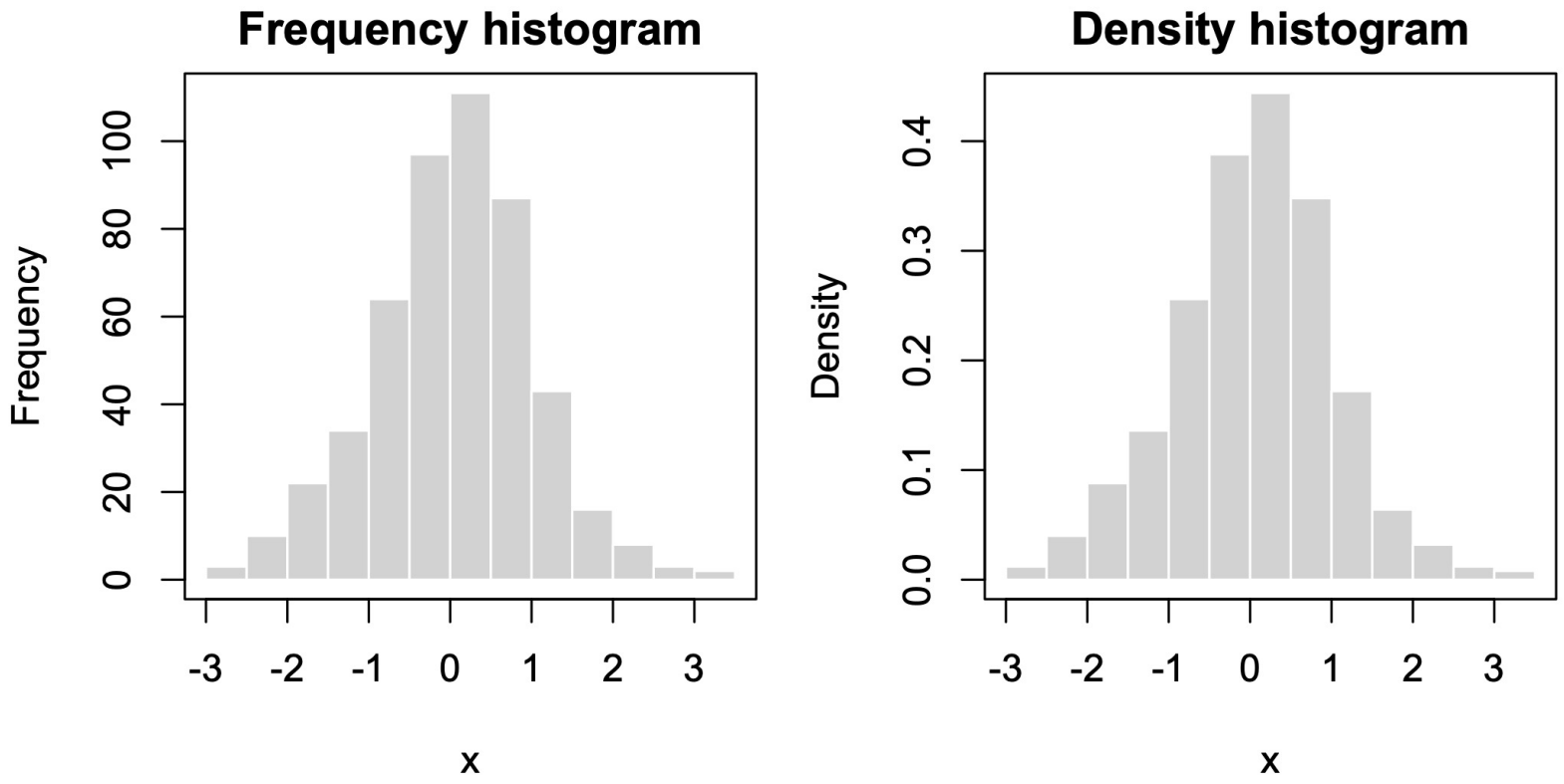
A **histogram** visualizes the empirical distribution of the data. To construct a histogram, we follow the steps below:

- divided the range of data into k bins, typically of equal width w;
- count the frequency of data within each bin f_i;
- draw a rectangle having the bin as its base. For frequency histogram, the height of the rectangle is equal to the frequency f_i; for a relative frequency histogram, or density histogram, the height of the rectangle is $f_i/(nw)$ so that the area of the rectangle is the relative frequency f_i/n.

Histogram

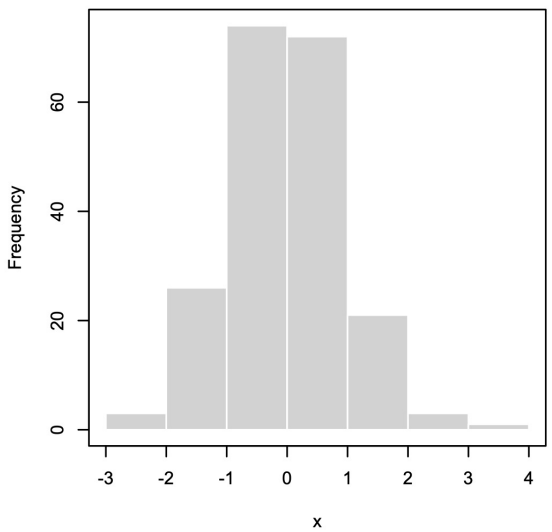
The frequency histogram and density histogram looks exactly the same. The only difference lies in the value of the y-axis.

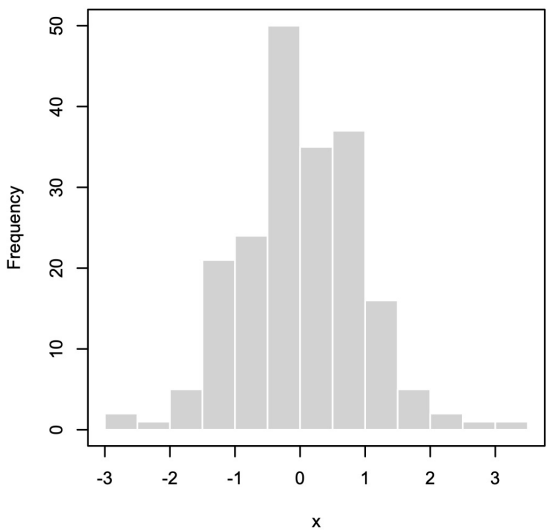
Density histogram is directly comparable to probability density function, and is thus more useful if you want to overlay PDF on top of the histogram.

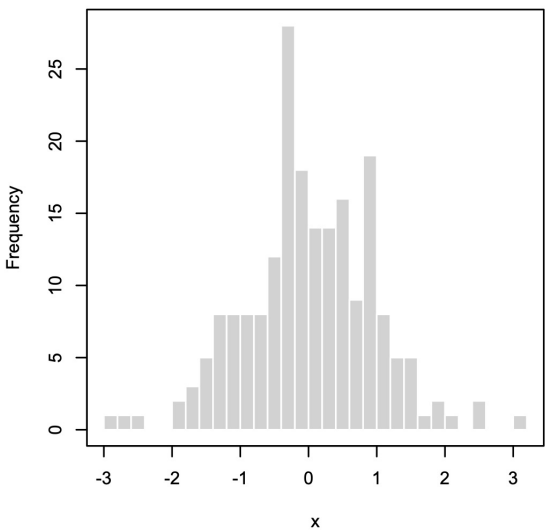


Histogram

There is no rule for how you choose the bins. But the width of the bins can change the look of the histogram. Thus, to comprehend the empirical distribution, it is often recommended that you use multiple choices of bins to visualize the data.



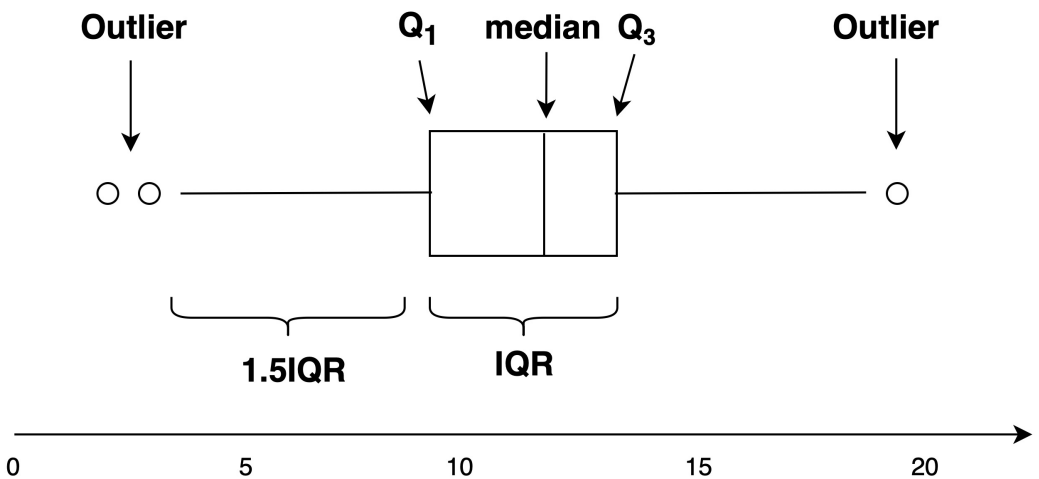




Box plot

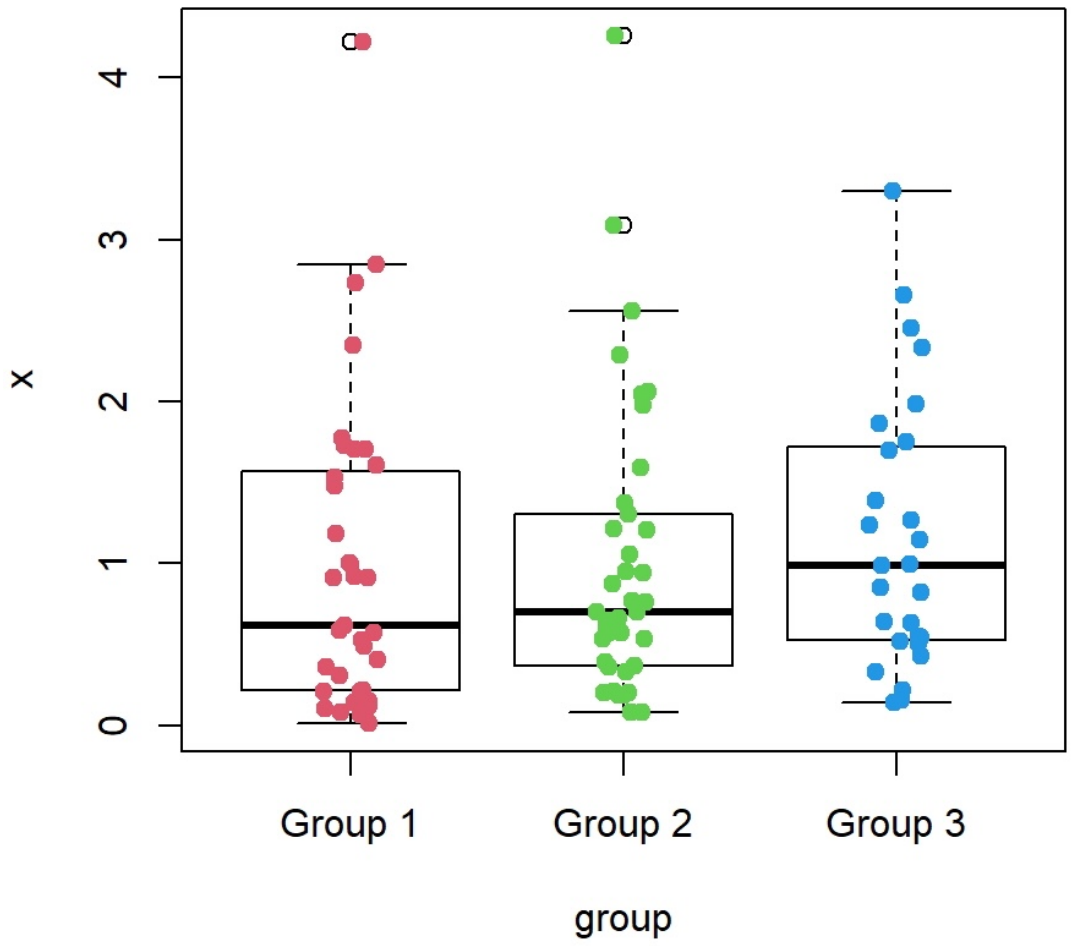
A box plot visualize the data distribution using a five number summary: minimum, maximum, the median, the first and third quartiles.

- a box drawn between Q_1 and Q_3 with a line inside indicating the median;
- Whiskers drawn to the largest observation smaller than 1.5IQR above Q_3 or the smallest observations larger than 1.5IQR below Q_1.
- dots indicating observations outside the boundaries of the whiskers, called outliers.



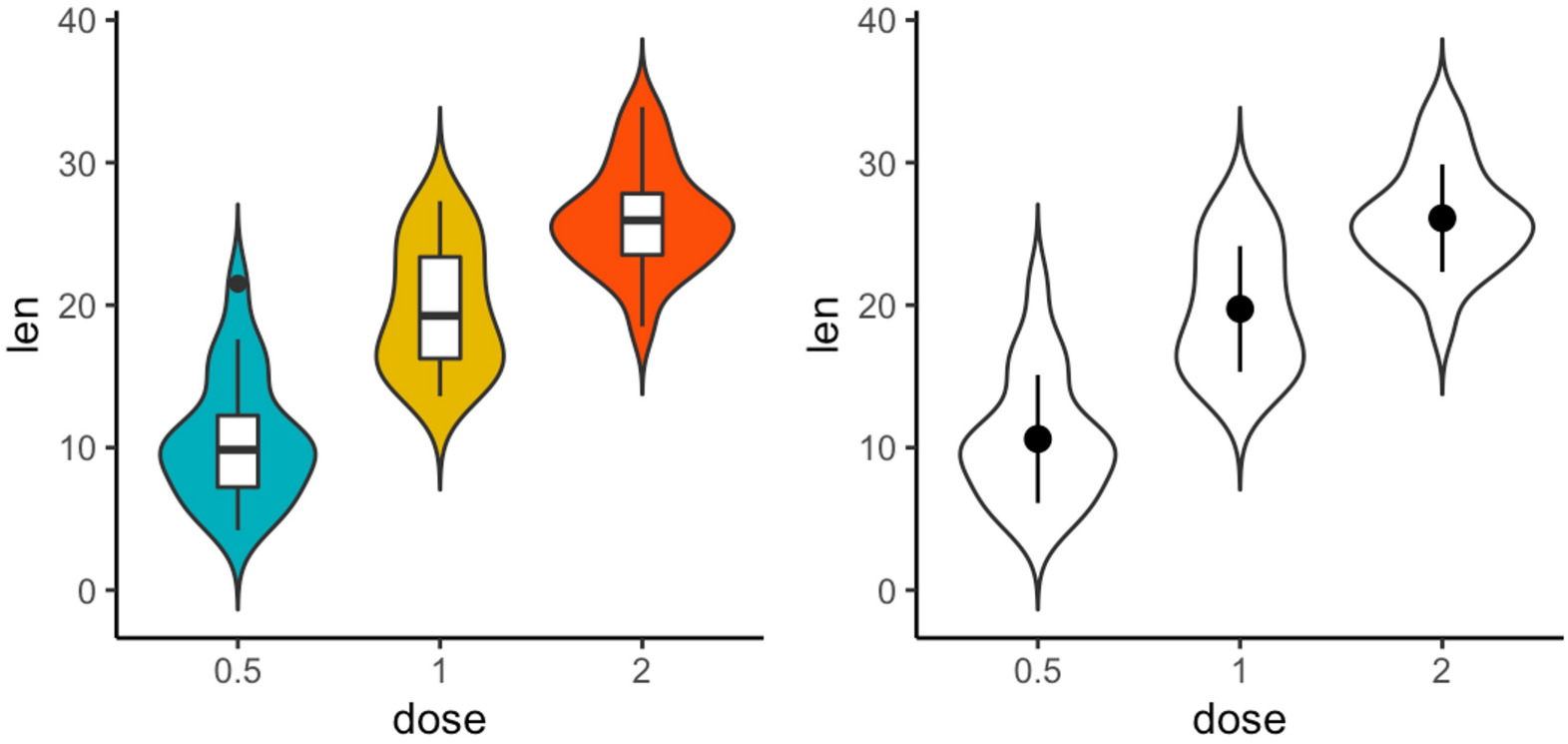
Box plot

A box plot displays a few summary statistics of the data. Raw data can be overlaid on a box plot to show the full distribution of the data.



Violin plot

The **violin plot** was proposed by Jerry Hintze and Ray Nelson in 1977 as a way to show more data than a box plot. It is essentially a box plot with an estimated probability density curve laid vertically on top of it.



Correlation coefficient

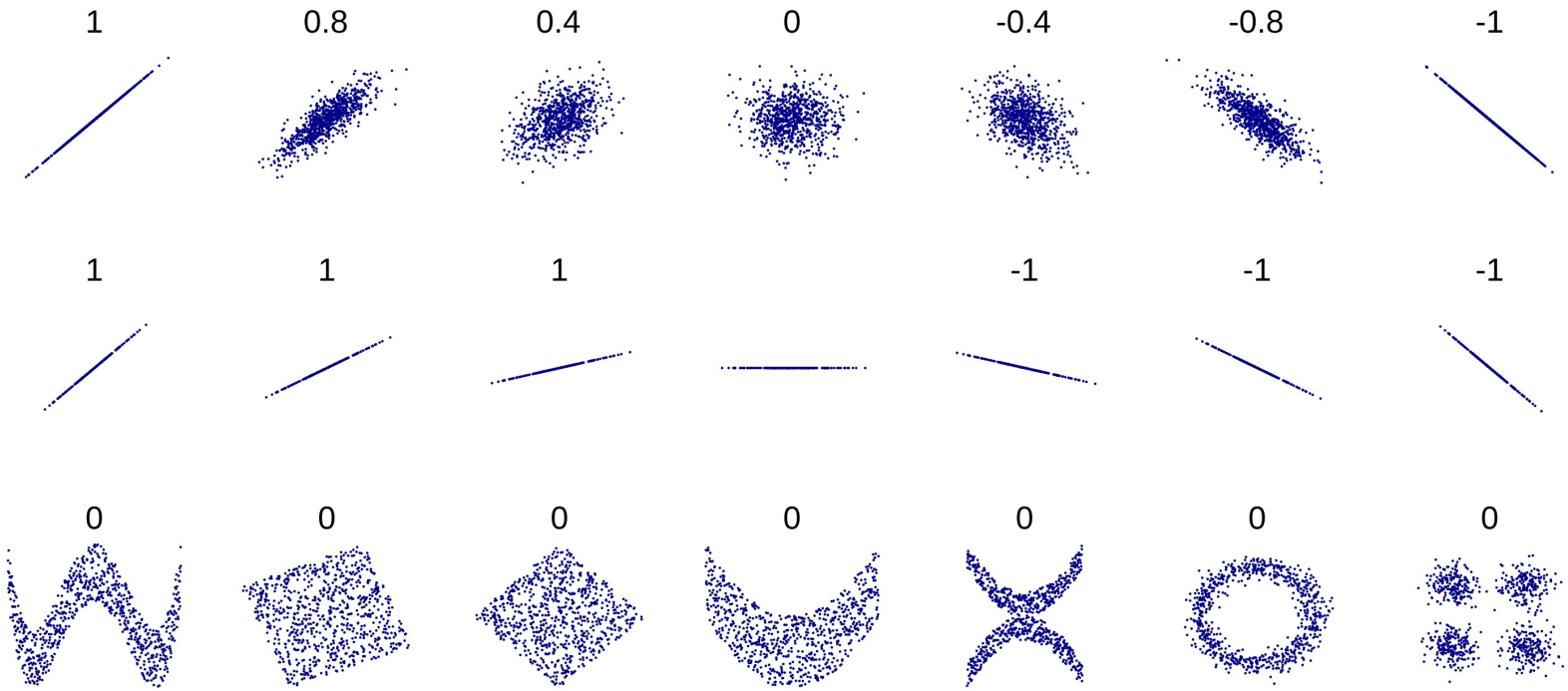
So far, we have discussed descriptive statistics and visualization for a univariate variable. For bivariate variables, a very commonly used descriptive statistic is the sample correlation coefficient.

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y}$$

where s_x and s_y are the sample standard deviations, and $\overline{x}$ and $\overline{y}$ are the sample means.

Correlation coefficient

The correlation coefficient reflects the noisiness and direction of a linear relationship, but not the slope of that relationship or the various aspects of nonlinear relationships.



Visualizing bivariate data

The type of figure used to visualize data primarily depends on data type:

- categorical vs categorical: stacked bar chart;
- numerical vs categorical: box plot or violin plot;
- numerical vs numerical: scatter plot

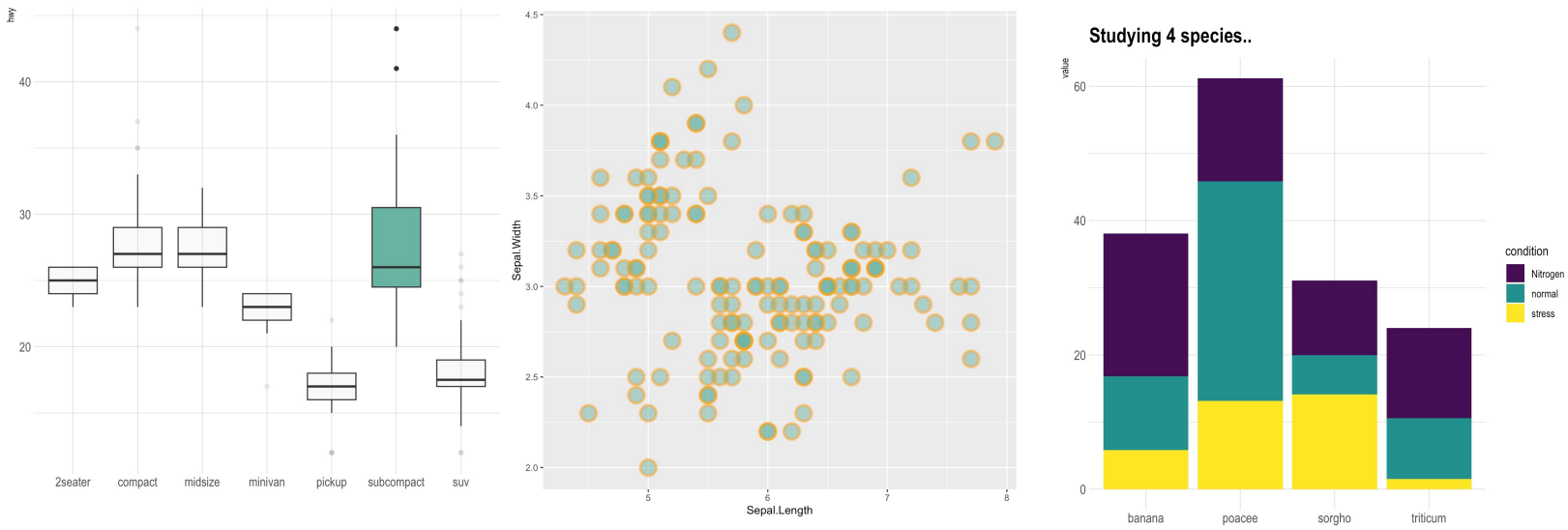
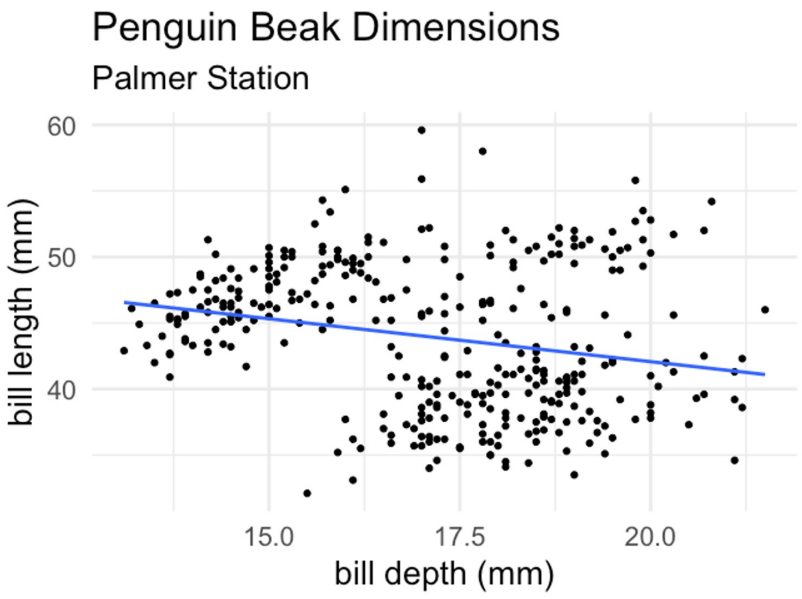
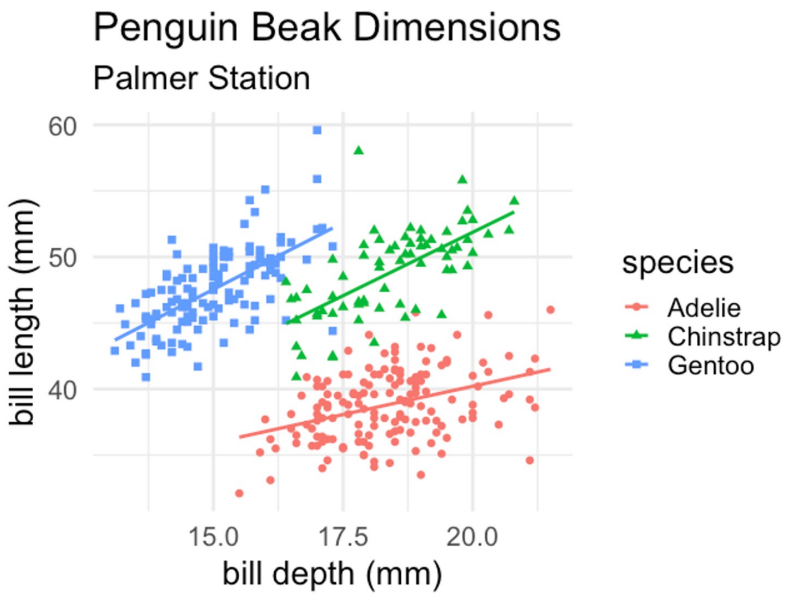


Illustration of different figure types of data visualization

Visualizing bivariate data

The data may have many hidden or nonobvious patterns in it. When the data can be divided into groups or categories, examining patterns within and across groups is often necessary.





Contrasting findings when data are analyzed in aggregate or in groups.