BIOS:4120 – Introduction to Biostatistics
Unit 2: Data Presentation

Knute D. Carter

Department of Biostatistics
The University of Iowa

August 28, 2025

Learning Objectives

At the end of this session, you should be able to:

- Describe, identify, and distinguish different types of data.
- Create frequency, relative frequency, and cumulative frequency tables.
- Construct the different types of graphs discussed if given a data set.
- Determine percentiles from a data set, and calculate the interquartile range.
- Identify if data are symmetric, left, or right skewed.

Overview

- Types of Data
- Frequency Tables
- Types of Graphs

Broad Distinctions

There are two broad distinctions we use to distinguish different types of data. We make a distinction between:

- *discrete* and *continuous* data;
- and also between *categorical* (qualitative) and *quantitative* (numerical) data.

As we will see there is overlap in these distinctions.

Categorical: Nominal Data

Nominal Data: The underlying variable assumes different categorical values, and these values have no inherent ordering.

Categorical: Dichotomous (or Binary) Data

Binary Data: Nominal data for which there are only two categories.

-
-
-
-
-

Categorical: Dichotomous (or Binary) Data

- Binary data is often coded as ‘0’ and ‘1’, though this is entirely arbitrary.
- Some statistical packages may require binary data to be coded numerically for some procedures.
- Even if one state is preferred over the other, binary data are still considered to be nominal.

Categorical: Ordinal Data

Ordinal Data: The underlying variable assumes different categorical values, and these values have an inherent ordering.

-
-
-
-

Categorical: Ordinal Data

- Any labeling of ordinal data only indicates the ordering of the categories.
- Suppose we used the labels, '0', '1', '2', and '3' for minor, moderate, severe, and fatal, respectively. This *does not* imply that a '2' is twice a '1'; or that the difference between a '1' and a '0' is the same as the difference between a '3' and a '2'.
- We could have used 'A', 'B', 'C', and 'D', the critical thing is that we can identify the inherent ordering.

Ranked Data

Ranked Data: The underlying variable assumes different numeric values or ordered categorical values, yet these values are replaced by an integer ranking that represents only relative position.

- The 5 most common types of cancer among white non-Hispanic American males
  - $1 =$ prostate
  - $2 =$ lung and bronchus
  - $3 =$ colon and rectum
  - $4 =$ urinary and bladder
  - $5 =$ Non-Hodgkin's lymphoma
- U.S. News and World Report rankings of the top 50 national public research universities.

Quantitative: Continuous Data

Continuous Data: The underlying variable assumes different numeric values, and the set of possible values consists of all real numbers over some interval or continuum.

-
-
-
-

Interval versus Ratio Level Data

- Within quantitative data one further distinction is sometimes made between *Interval Data* and *Ratio Data*.
- *Interval Data* has no natural zero, whereas *Ratio Data* does.
- The almost singular example of interval data is that of temperature scales.

Quantitative: Discrete Data

Discrete (count) Data: The underlying variable assumes different numeric values, yet the set of possible values is restricted to a specific list of numbers (often integers).

-
-
-

Here *both* order and magnitude are important.

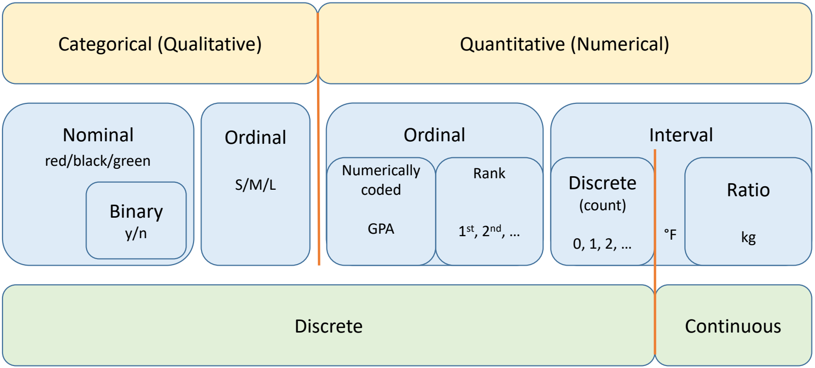
Further Notes on Types of Data

- With nominal and ordinal data, the underlying variable is non-numeric, yet may be coded as numeric.
- With discrete (interval) and continuous data, the underlying variable is inherently numeric.
- With ranked data, the underlying variable may be numeric (e.g., incidence rates of cancer), yet the magnitudes of the values are discarded and only the relative positions are used.
- A measurement on a variable is often called an *observation*.

Further Notes on Types of Data

- The measurement scale is important when deciding how to analyze data. For example, don't report a sample mean of a nominal variable!
- Sometimes ordinal data are converted to scores for analysis. This may be controversial.
- Even when we have an underlying continuous variable, we often measure and/or present it according to it discrete values.

Schematic of Types of Data



Frequency Tables

- Tables can provide a convenient method for summarizing and presenting the distribution of data.
- Suppose we had the disease stage from 50 cancer patients. We could summarize the data in the following *frequency table*:

Stage	Frequency
I	8
II	15
III	11
IV	16
Total	50

- The *frequency* is simply the number of observations in each category.

Frequency Tables

- *Relative Frequency*: The proportion (or percentage) of observations in each category.

Stage	Frequency	Relative Frequency
I	8	0.16
II	15	0.30
III	11	0.22
IV	16	0.32
Total	50	1.00

Frequency Tables

- For continuous data we may only have one observation at any given value, so we need to do something a little different.

- Consider a sample of 25 cancer patients with the following disease free survival (DFS) times (in months):

```
 1,  2,  3,  5,  7,  8,  8,  9, 10, 10, 11, 11, 12,
12, 13, 14, 15, 17, 18, 19, 21, 22, 34, 35, 39.
```

  - The DFS time of a treated cancer patient is defined as the length of elapsed time between the time at which the patient goes into remission and the time at which the patient relapses.

- To summarize this data, we form a set of non-overlapping intervals into which all of the data can be grouped.

Distribution of DFS Times in Months

Interval numbers i	Class interval	Frequency f_i	Relative frequency p_i
1	0–4	3	0.12
2	5–9	5	0.20
3	10–14	8	0.32
4	15–19	4	0.16
5	20–24	2	0.08
6	25–29	0	0.00
7	30–34	1	0.04
8	35–39	2	0.08

- *Frequency*: The number of observations in an interval.
- *Relative Frequency*: The proportion (or percentage) of observations in an interval.

Distribution of DFS Times in Months

Interval numbers i	Class interval	Frequency f_i	Relative frequency p_i	Cumulative relative frequency
1	0–4	3	0.12	0.12
2	5–9	5	0.20	0.32
3	10–14	8	0.32	0.64
4	15–19	4	0.16	0.80
5	20–24	2	0.08	0.88
6	25–29	0	0.00	0.88
7	30–34	1	0.04	0.92
8	35–39	2	0.08	1.00

- *Cumulative Relative Frequency*:
  The proportion (or percentage) of observations that are less than or equal to the upper endpoint of the interval.

Graphs

The pattern of variation of measurements on a variable is referred to as the *Distribution* of the variable.

There are two approaches to summarizing a distribution:

- Graphical (Visual)
- Arithmetic

Types of Graphs

We will review the following types of graphs:

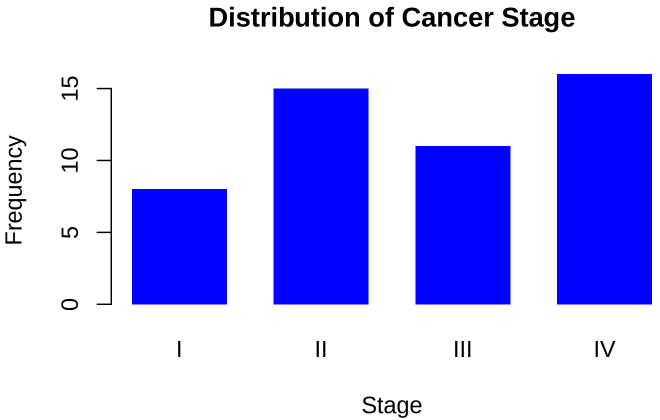
- Bar Charts
- Histograms
- Frequency Polygons
- Stem-and-Leaf Diagrams
- One-Way Scatter Plots
- Box Plots
- Two-Way Scatter Plots

Bar Charts

- Bar charts are very common, and are used to show the relative frequency observed in different (nominal or ordinal) categories.
- They are characterized by having spaces between the bars. (For nominal data, the order of the bars is irrelevant).
- The measurement scale on the height of the bar, either vertically or horizontally, could be either absolute frequencies or relative frequencies.
- **IMPORTANT**: The bars should be of equal width. Why?

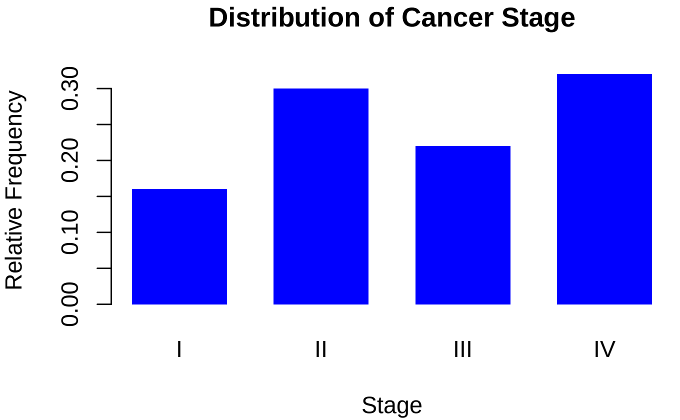
Bar Charts

The following represents a bar chart constructed from the cancer stage data:



Bar Charts

The following represents a bar chart constructed from the cancer stage data:



Histograms

- A *Histogram* is a graphical representation of a frequency or relative frequency distribution for discrete or continuous variables.
- They are similar to bar graphs, except the bars touch each other.
- It is the **area** of the bar that reflects the relative proportion, rather than the height of the bar.
- If the bars are the same width, the height can be the frequency and the area will be appropriate too.

Constructing a Histogram

1. Construct the (relative) frequency distribution using non-overlapping intervals of equal width.
2. Create horizontal and vertical axes for the graph.
   - Label the vertical axis so as to accommodate all of the numbers in the (relative) frequency distribution.
   - The vertical scale should begin at zero.
   - Label the horizontal axis with the endpoints for the intervals, or with the interval numbers.
3. Construct a rectangle over each interval, with the height representing the frequency (relative frequency) for that interval.

Histogram Example

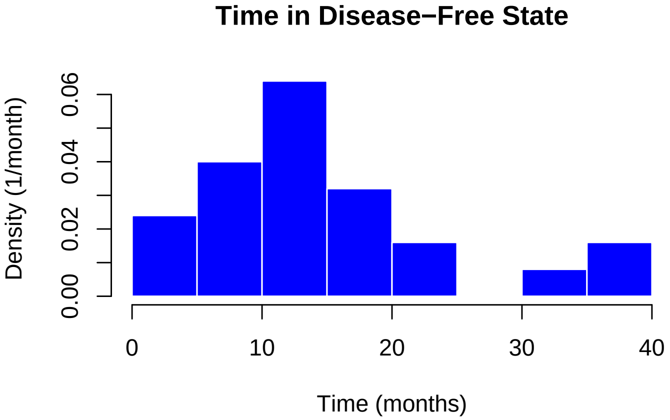
The following represents a histogram for the DFS time data.

Histogram Example

The following represents a histogram for the DFS time data.

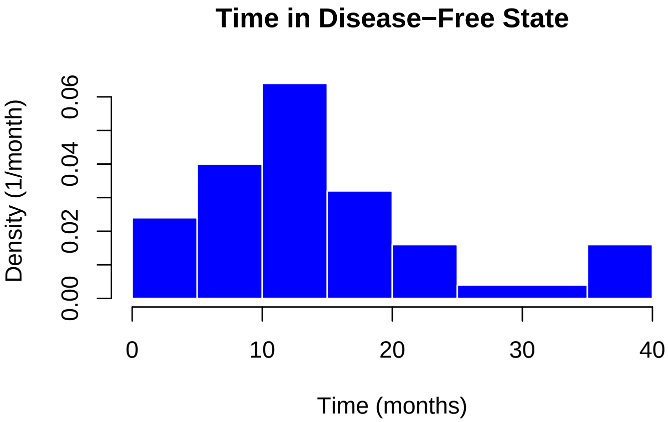
Histogram Example

The following represents a histogram for the DFS time data.



Histogram Example

The following represents a histogram for the DFS time data.

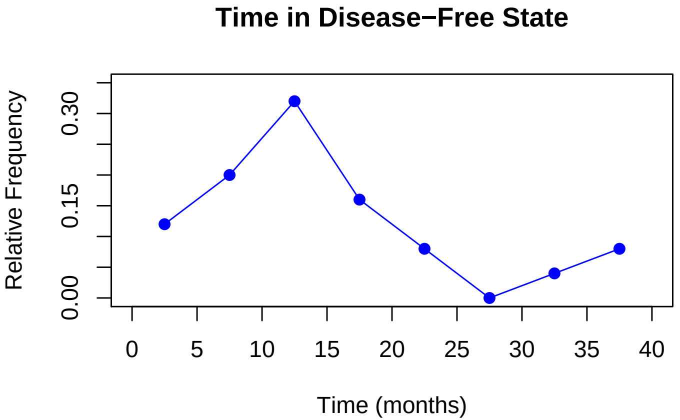


Frequency Polygons

- Like a histogram, a *Frequency Polygon* is a graphical representation of a relative frequency distribution.
- A *Cumulative Frequency Polygon* is a graphical representation of a cumulative frequency distribution.
- A frequency polygon uses the same axes as a histogram. It is constructed by placing a point over the center of each interval such that the height of the point corresponds to the relative frequency associated with the interval. The points are then connected.

Frequency Polygon Example

The following represents a frequency polygon constructed from the relative frequency distribution of the DFS data:

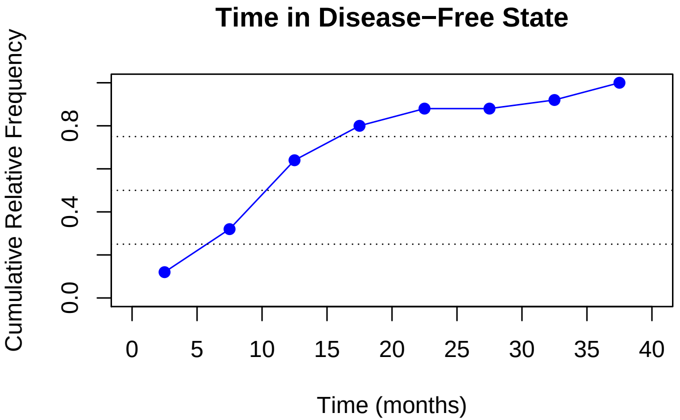


Frequency Polygons

- Frequency polygons are sometimes preferred to histograms for comparing two or more distributions, since they can be easily superimposed.
- Cumulative frequency polygons are useful for identifying the *percentiles* of a data set.
- The p^{th} *Percentile* of a data set $(0 \leq p \leq 100)$ is a value which exceeds about $p\%$ of the observations, and is exceeded by about $(100 - p)\%$ of the observations.
- The *Median* of a data set corresponds to the 50^{th} percentile.
- The *First*, *Second*, and *Third Quartiles* of a data set correspond to the 25^{th}, 50^{th}, and 75^{th} percentiles, respectively.

Cumulative Frequency Polygon Example

The following represents a cumulative frequency polygon from the relative frequency distribution of the DFS data:



Stem-and-Leaf Diagrams

- *Stem-and-Leaf Diagrams* are like histograms (on their side) built out of the numbers themselves.
- Easier and quicker to construct for small data sets by hand.

Stem-and-Leaf Diagram Example

- Suppose we had the ages of 20 patients as follows:

```
58, 56, 50, 48, 66, 43, 71, 60, 78, 54,
57, 45, 67, 53, 36, 70, 64, 65, 56, 57.
```

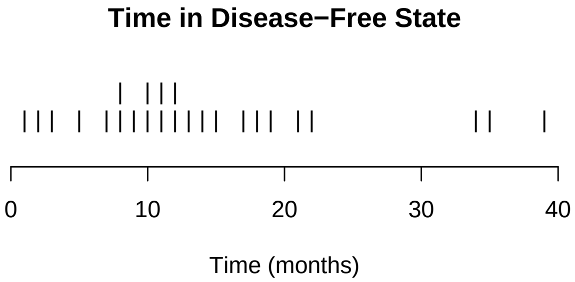
- A stem-and-leaf diagram for these data would look like:

```
3 | 6
4 | 8 3 5
5 | 8 6 0 4 7 3 6 7
6 | 6 0 7 4 5
7 | 1 8 0
```

- The *stem* is the tens digit, and the *leaf* is the units digit.
- Other modifications exist of stem-and-leaf diagrams.

One-Way Scatter Plot

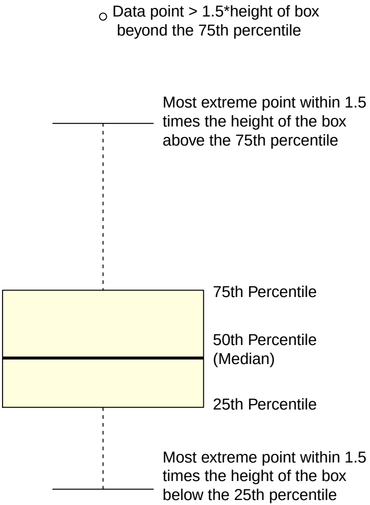
- A one-way scatter plot simply uses tick marks on a horizontal scale to indicate where the data points fell.



Box Plots (a.k.a. Box-and-Whisker Plots)

- *Box Plots* visually depict a lot of information about the whole set of numbers (percentiles, skewness, outliers).
- *Outliers* are atypical values in a data set.
- Like a one-way scatter plot, they only use one axis, but they provide a lot more information.

Box Plot Construction



Percentiles

Here is an intuitive definition of a *percentile* in a sample.

The p^{th} percentile, V_p, is a number such that:

1. $p\%$ of the observations are less than V_p, and
2. $(100 - p)\%$ of the observations are greater than V_p.

Percentiles, an Example

- The 10^{th} percentile is a number that is greater than or equal to 10% of the data, but smaller than the other 90%.
- If you had the ten numbers

  `2, 3, 5, 5, 6, 7, 10, 23, 26, 28`

  as your sample data, then the 10^{th} percentile would be any number between 2 (including 2) and less than 3.

Percentiles: An Example

- Difficulties arise with this definition when (1) the sample size is small, (2) there are tied values, and (3) the percentiles are not unique.
- Suppose you had

  `2, 2, 5, 5, 6, 7, 10, 23, 26, 28`
- Now what would the 10^{th} percentile be?
- There is no single definition for calculating percentiles.
- In fact R provides 9 different methods to choose from!

Percentiles: One Formal Definition

1. Rank the observations from smallest to largest ($1 =$ smallest, $n =$ largest)
2. If $np/100$ is not an integer, V_p is the k^{th} observation, where k is the smallest integer greater than $np/100$.
3. If $np/100$ is an integer, V_p is the average of the $(np/100)^{\text{th}}$ and $(np/100 + 1)^{\text{th}}$ observations.

Percentiles: An Example

Data: `2, 3, 5, 5, 6, 7, 10, 23, 26, 28.`

- $n = 10$, and we want the $p = 10^{\text{th}}$ percentile.
- Now what would the 10^{th} percentile be?
- Since $np/100 = (10 \times 10)/100 = 1$ an integer,
  then the 10^{th} percentile is the average of the
  1^{st} $(= np/100)$, and 2^{nd} $(= np/100{+}1)$, observations.
- And so the $p = 10^{\text{th}}$ percentile $= (2 + 3)/2 = 2.5$.

Data: `2, 2, 5, 5, 6, 7, 10, 23, 26, 28.`

- By the same reasoning, the $p = 10^{\text{th}}$ percentile would be the average of the 1^{st} and 2^{nd} observations, which is $(2 + 2)/2 = 2$.

Percentiles: An Example

Now increase the sample size by one.

Data: `1, 2, 2, 5, 5, 6, 7, 10, 23, 26, 28`. ($n = 11$)

- Since $np/100 = (11 \times 10)/100 = 1.1$, not an integer, then the 10^{th} percentile is now defined as the the k^{th} observation, where k is the smallest integer greater than 1.1 ($= np/100$).
- Thus the $p = 10^{\text{th}}$ percentile becomes the 2^{nd} observation, which is 2.

Quartiles

The *Quartiles* of the data are defined as:

- $Q_1 = 25^{\text{th}}$ percentile $= 1^{\text{st}}$ or lower quartile
- $Q_2 = 50^{\text{th}}$ percentile $= 2^{\text{nd}}$ quartile or median $= M$
- $Q_3 = 75^{\text{th}}$ percentile $= 3^{\text{rd}}$ or upper quartile

The *Interquartile Range (IQR)* is defined as $Q_3 - Q_1$.

Steps to Construct a Box Plot

1. Determine the three quartiles: Q_1, M, Q_3.
2. Determine the *Interquartile Range*:

$$IQR = Q_3 - Q_1$$

3. Construct a reference axis. Label the axis so as to accommodate all of the numbers in the data set.
4. Draw a segmented box next to the reference axis to represent the quartiles.

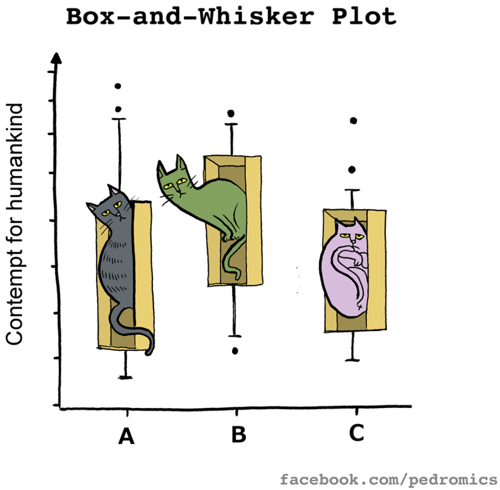
Steps to Construct a Box Plot (continued)

5 From the upper and lower quartiles, draw lines (*whiskers*) that extend out to the last observation which is no further than $1.5 \times IQR$ away from the box.

Thus, the line cannot be any longer than $1.5 \times IQR$.

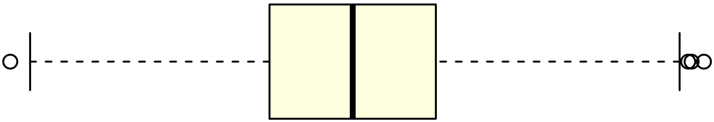
6 Any observations that lie beyond the end of a whisker are regarded as outliers, and are marked with a special symbol, (e.g., a circle or a star).

What's Wrong with these Box Plots?



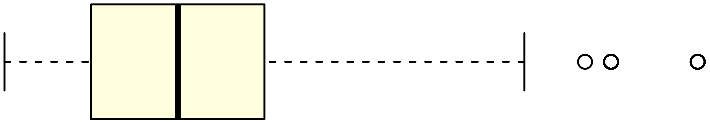
Identifying Skewed Data from a Box Plot

- A distribution is *Symmetric* if the left side of the histogram is a mirror image of the right side.
- Such a distribution may have a box plot such as:



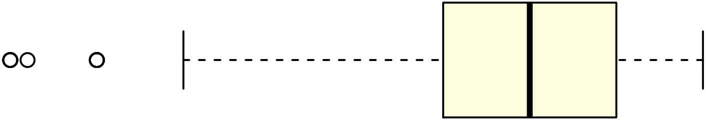
Identifying Skewed Data from a Box Plot

- A distribution is *Skewed Right* if the majority of area for the histogram is on the left, and the area gradually trails off on the right.
- Skewed right (long right whisker, toward large numbers):



Identifying Skewed Data from a Box Plot

- A distribution is *Skewed Left* if the majority of area for the histogram is on the right, and the area gradually trails off on the left.
- Skewed left (long left whisker, toward small numbers):



Two-Way Scatter Plots

- A *Two-Way Scatter Plot* is simply a two-dimensional plot of data pairs.
- It used to visualize the relationship between two variables.
- Examples:
  - A patient's BMI (body mass index) versus his/her total cholesterol
  - A patient's systolic blood pressure versus his/her diastolic blood pressure

Two-Way Scatter Plot Example

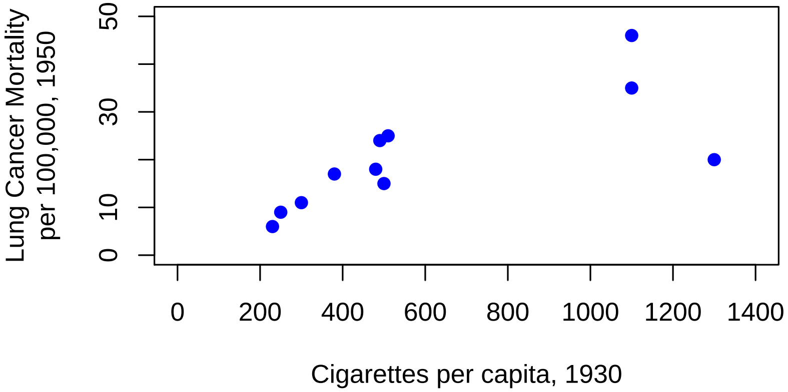
- Consider the data presented in the following table: per capita cigarette consumption in 1930, and lung cancer cases per 100,000 in 1950.

Country	Cigarettes (1930)	Lung Cancer (1950)
USA	1300	20
Great Britain	1100	46
Finland	1100	35
Switzerland	510	25
Canada	500	15
Holland	490	24
Australia	480	18
Denmark	380	17
Sweden	300	11
Norway	250	9
Iceland	230	6

Doll, R. (1955). Etiology of lung cancer.
Advances Cancer Research, *3*, 1–50.

Two-Way Scatter Plot Example

- We want to explore if there is a relationship (trend) between cigarette consumption and lung cancer.



Learning Objectives

At the end of this session, you should be able to:

- Describe, identify, and distinguish different types of data.
- Create frequency, relative frequency, and cumulative frequency tables.
- Construct the different types of graphs discussed if given a data set.
- Determine percentiles from a data set, and calculate the interquartile range.
- Identify if data are symmetric, left, or right skewed.