

2. Data Presentation

2.1. Types of Numerical Data

Statistical methods are often meant to be applied to a particular type of data. Thus, before we learn how to use methods to describe and to make inference from our data, we should learn something about the types of data we may encounter.

2.1.1. Nominal data

Think: “Name” for Nominal. That is, nominal data are data consisting of **names or labels**.

E.g. Blood Type: A, B, AB, O

E.g. Gender: Female, Male

E.g. Ethnicity: Black, Caucasian, Hispanic, Native American, Oriental

E.g. Disease Status: Infected, Not Infected

Note that the above examples do not suggest any natural ordering to the data. Often, nominal data are coded with numbers.

E.g. 1 for Female, 0 for Male--But, this does not necessarily mean that a Male (0) is less than a Female (1)!

Dichotomous or *binary* data is nominal data with 2 categories (E.g. Disease Status, Yes/No, 0/1).

2.1.2. Ordinal data

Ordinal data is like nominal data, but with some **natural ordering**.

E.g. Injury Severity: 1-Fatal, 2-Severe, 3-Moderate, 4-Minor

E.g. Disease Risk: 1-High, 2-Moderate, 3-Low

E.g. Headache Relief: 1-None, 2-Mild, 3-Moderate, 4-Complete Relief

Note the natural ordering: Fatal is certainly more severe than Moderate. **But, the numbers are merely codes. They usually don't have much meaning.** For example, 2 is not twice as much as 1 in this case.

2.1.3. Ranked Data

Ranked data is numerical data that has been ordered according to the magnitude of the data values and then assigned a "rank" number (high to low or vice versa).

E.g. Table 2.3 (POB)

Rank	Cause of Death	Total Deaths
1	Diseases of the heart	717,706
2	Malignant neoplasms	520,578
3	Cerebrovascular diseases	143,769
4	Chronic obstructive pulmonary diseases	91,938
5	Accidents and adverse conditions	86,777
6	Pneumonia and influenza	75,719
7	Diabetes mellitus	50,067
8	HIV infection	33,566
9	Suicide	30,484
10	Homicide and legal intervention	25,488

Note that the rank values imply order, but **differences and ratios of ranks are likely meaningless**. For example, suicide (9) does not account for 3 times the number of deaths as cerebrovascular diseases (3).

2.1.4. Discrete Data

Nominal and ordinal data are not typically considered as numerical data (though they may have numerical codes). Discrete data are **numbers with order and magnitude**, but with no possibility of values between adjacent data values. **Count data (non-negative integers)** are a very common type of discrete data (0,1,2,3,...).

E.g. Number of births

E.g. Number of deaths (Can't have 3.5 deaths)

E.g. Number of accidents

Note that the **difference**, **ratio** and other **arithmetic operations** on discrete data usually make sense (although the ratio or average may not be an integer).

2.1.5. Continuous Data

Like discrete data, continuous data have meaningful order and magnitude, but also take on all possible values in a continuum (0 to 1, all positive values, any real number). Measurement data are frequently continuous data.

E.g. Height (180.5 cm), Weight (75 kg), Temperature, Cholesterol Levels, Age (?), Pressure (could be negative)

Note that sometimes it may not make sense to take the ratio of numerical data. For example, does 80 degrees Fahrenheit indicate twice as much heat as 40 degrees Fahrenheit?

2.2. Tables

Consider the raw data in Table 2.1 page 8 of POB. What can you say about AIDS patients with Kaposi's sarcoma from looking at the raw data? Hmmm. We need methods to summarize the data in order to make meaningful statements about it. Both tables and graphs are used to summarize data. In fact, it can be argued that these should always be used in the first steps of analyzing data. What you lose in detail will likely be more than compensated by a gain in understanding.

Frequency (counts), Relative Frequency (proportion or percent), Cumulative Frequency (counts, proportions, or percent) and Cumulative Relative Frequency are often included in a table for data in categories. These tables (and graphs in next section) are sometimes referred to as frequency distributions for obvious reasons.

E.g. Serum Cholesterol Levels for 1067 U.S. males, aged 25 to 34 years, 1976-1980 (see POB Table 2.6).

Colesterol Level (mg/100 ml)	Frequency	Cumulative Frequency	Relative Frequency (%)	Cumulative Relative Frequency (%)
80-119	13	13	1.2	1.2
120-159	150	163	14.1	15.3
166-199	442	605	41.4	56.7
200-239	299	904	28.0	84.7
240-279	115	1019	10.8	95.5
280-319	34	1053	3.2	98.7
320-359	9	1062	0.8	99.5
360-399	5	1067	0.5	100.0
Total	1067	1067	100.0	100.0

Verify some of the values in the table:

If two groups have different numbers of observations, then **comparisons across groups** must somehow be adjusted for the different number of observations. For example, in the table below, it wouldn't make much sense to compare the frequency (as counts; or CF) of men age 25-34 to men age 55-64 at a particular cholesterol category if the total number of men in each age group differs; your conclusions would depend on the totals! You want a measure that does not depend on totals. Relative frequency is often used to make such comparisons.

E.g. Similar to previous example (see POB Table 2.8)

Cholesterol Level (mg/100 ml)	Ages 25-34			Ages 55-64		
	Frequency	Relative Frequency (%)	Cumulative Relative Frequency (%)	Frequency	Relative Frequency (%)	Cumulative Relative Frequency (%)
80-119	13	1.2	1.2	5	0.4	0.4
120-159	150	14.1	15.3	48	3.9	4.3
166-199	442	41.4	56.7	265	21.6	25.9
200-239	299	28.0	84.7	458	37.3	63.2
240-279	115	10.8	95.5	281	22.9	86.1
280-319	34	3.2	98.7	128	10.4	96.6
320-359	9	0.8	99.5	35	2.9	99.4
360-399	5	0.5	100.0	7	0.6	100.0
Total	1067	100.0	100.0	1227	100.0	100.0

2.3. Graphs

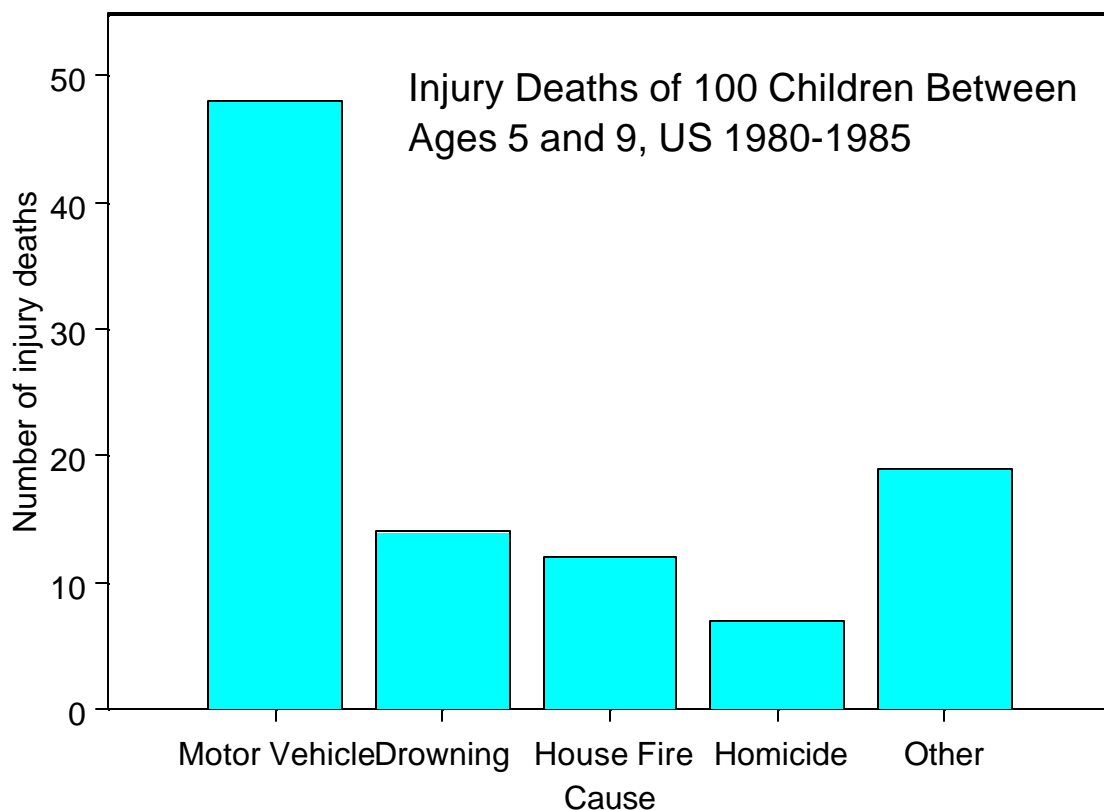
Graphs are often easier to interpret than tables, perhaps at the expense of detail. A variety of graphs are used depending on the type of data. We'll look at a few basic types. You can be creative and go beyond those graphs (and tables) we'll cover, but you should always have a **clear purpose** for a graph and it should always be **simple to understand** and **self-explanatory**.

2.3.1. Bar charts

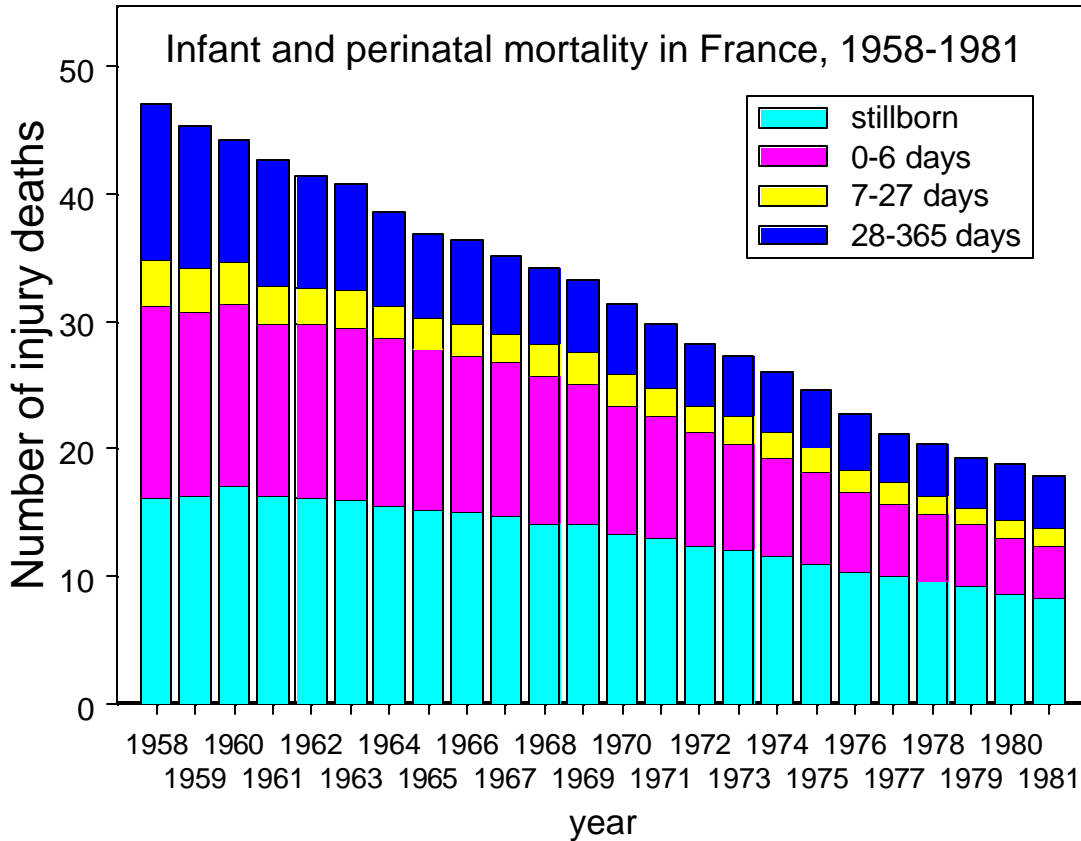
Bar charts are used to display the frequency distribution of categorical data (nominal or ordinal).

- Bars should be **equal width**
- Bars should be **separated** so as not to imply continuity of categories

E.g. Below is an example of a simple bar chart created with S-Plus (see also Figure 2.13 page 27 of POB).



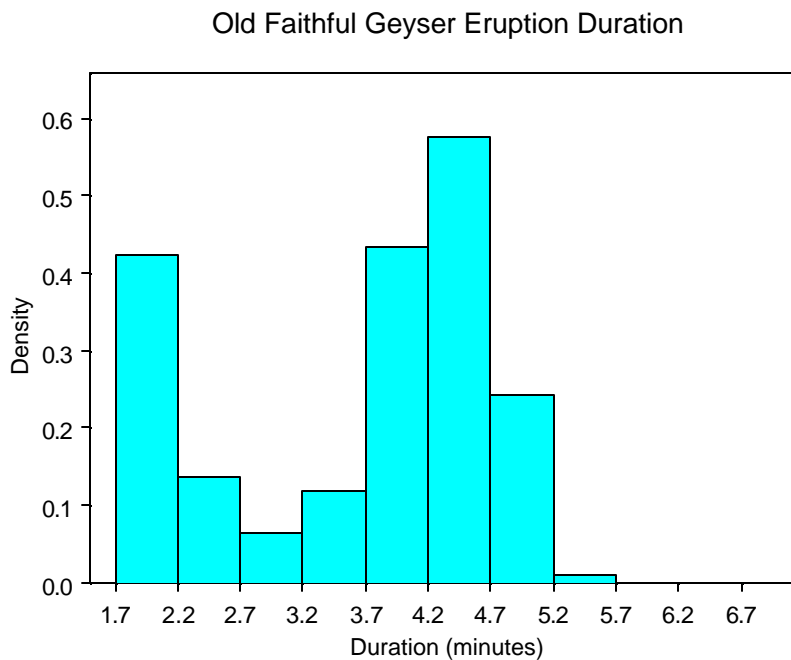
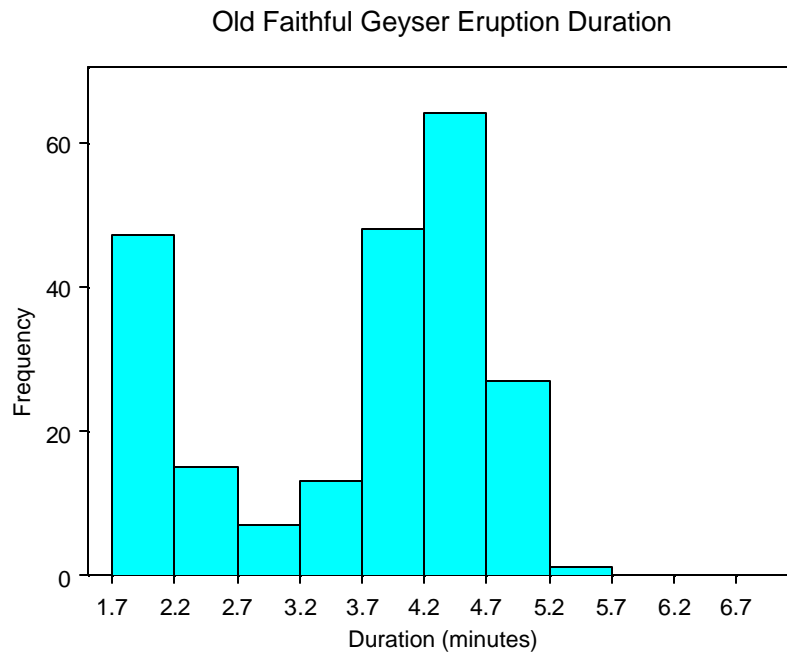
E.g. You will get a chance to create the **stacked bar chart** below in lab using S-Plus to show the frequency distribution of **2 or more subgroups simultaneously**



2.3.2. Histograms

Histograms depict **frequency** or **relative frequency** distributions for discrete or continuous data. They are perhaps the most frequently used graphical summary for quantitative data. They depict the overall distribution shape:

- unimodal, bimodal, or multi-modal
- bell-shaped, left-skewed, right skewed
- range or spread of data
- allow finding of proportions
- Bars are **connected** (implies continuity for continuous data)
- *Strictly speaking* the **area** of each histogram bar is **equal** to the **proportion** of observations falling in an interval:
 - (area of bar) = (intvl. ht * intvl. width) = (rel. freq.)
- Thus,
 - (intvl. ht) = (rel. freq.) / (intvl. width)
- In other words,
 - (intvl. ht) = (intvl. count/total count)/(intvl. width)
- If the interval width is chosen to be constant across intervals (almost always), then
 - (intvl. ht) = (some constant)*(rel.freq) or
 - (intvl. ht) = (some other constant)* freq.
- Thus, we *most often* see **height** as **freq.** or **rel. freq.** since the constant multiplier doesn't do anything to change the overall shape of the histogram.



Remember that a histogram should give you an overall impression of your data. You will often need to experiment with the number intervals and their width. Check out this applet that creates a histogram of the duration of eruptions of the Old Faithful geyser:

<http://www.stat.sc.edu/%7Ewest/javahtml/Histogram.html>

Experiment with the applet until you get a “good-shaped” distribution. **How does the bin width affect conclusions?** Notice the vertical axis on the applet is neither frequency nor relative frequency. The applet is making the area of the bar equal to relative frequency, not the height (except when bin width = 1, of course). In this case the height is sometimes called **density**.

2.3.3. Frequency polygons

- Plot **frequencies** or relative **frequencies** versus midpoint of intervals from a frequency distribution table (basically, connect the tops of histogram bars at their mid-point).
- Convenient for overlaying 2 or more distributions.
- **Cumulative relative frequency polygon** can be used to find **percentiles** (values for which a certain percent of the data fall at or below that value).

2.3.4. One-way scatter plots (see POB).

2.3.5. Box Plots

- Based on **quartiles** of a distribution
 - **25th** p-tile is the **1st** quartile (**Q1**); 25% of the data fall at or below this value and 75% fall at or above
 - **50th** p-tile is the **2nd** quartile or median (**Q2**); 50% of the data values fall at or below this value and 50% at or above
 - **75th** p-tile is the **3rd** quartile (**Q3**); 75% of the data fall at or below this value and 25% at or above
- Make a box from Q1 and Q3 and draw line at Q2. Extend “whiskers” out to values that are no more than 1.5 times the length of the box (i.e. $1.5 \cdot (Q3 - Q1)$) (these are called **adjacent values** in POB). Note that **Q3-Q1 is called the interquartile range (IQR)** Denote values more extreme than the whiskers (**outliers**) with a line or dot or star or whatever.
- Indicates symmetry or skewness
- Indicates “outliers”

E.g. (Note that this is a revision of an earlier (and incorrect) procedure for calculating quartiles. Please correct your notes. This version is correct.)

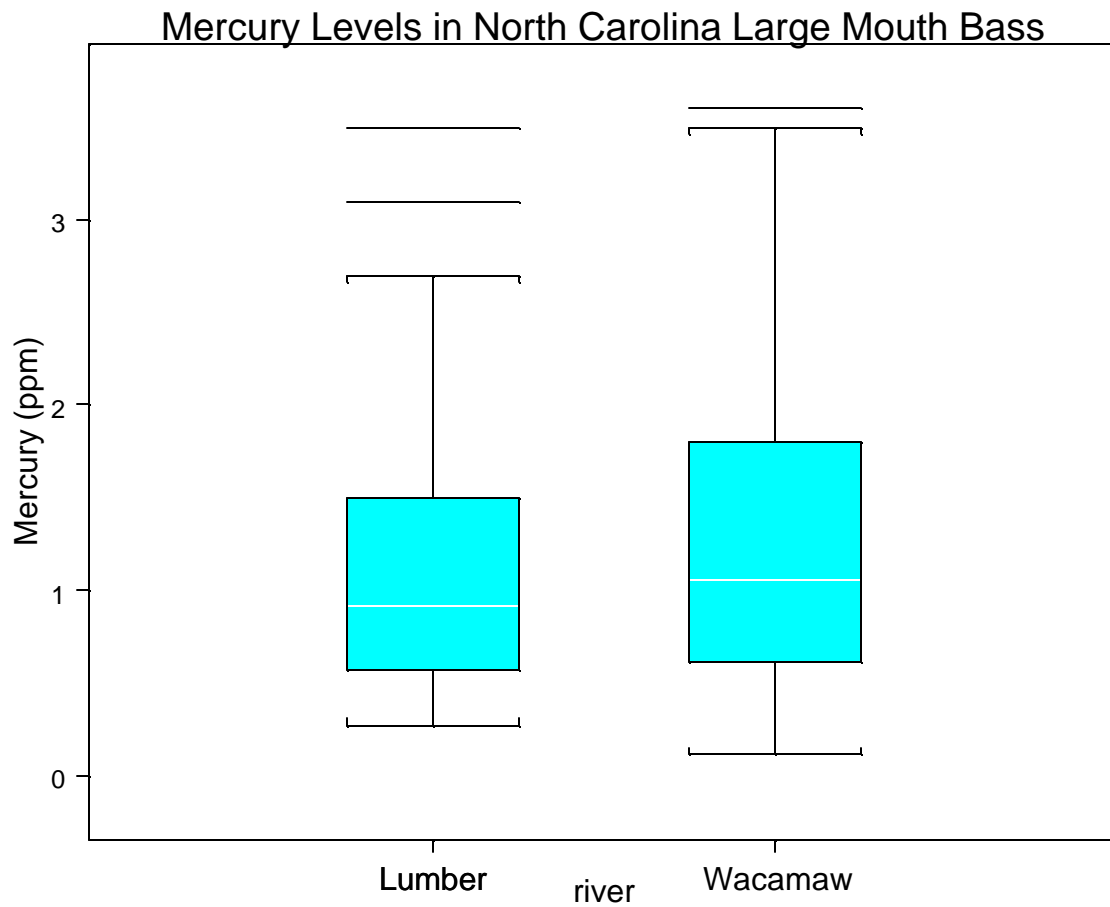
Some data: 2 3 9 5 6 4 8 7 6 4 9 7 3 6 4

1) Order the data: 2 3 3 4 4 4 5 6 6 6 7 7 8 9 9

2) Odd number of data so **Q2** is middle value: ____ (if even take average of middle 2 values)

- 3) **Q1** is the 25th percentile. Looking ahead to chapter 3, we follow the rules for calculating percentiles:
- $n*k/100 = 15*25/100=3.75$ ($k=25$ here), so we round up to 4 (according to rules for calculating percentiles in chapter 3) and take the 4th ordered data value: _____
- 4) **Q3** is the 75th percentile. Again, using the rules for calculating percentiles in chapter 3 we have:
- $n*k/100 = 15*75/100 = 11.25$ ($k=75$ here), so we round up to 12 and use the 12th ordered data value: _____ (position is “mirror image” of Q1)
- 5) $1.5*(Q3-Q1) = 1.5*_____ = _____$
- 6) No values are less than $Q1-4.5$ or greater than $Q3+4.5$ so draw whiskers to min (2) and max (9)
- 7) Draw the plot:

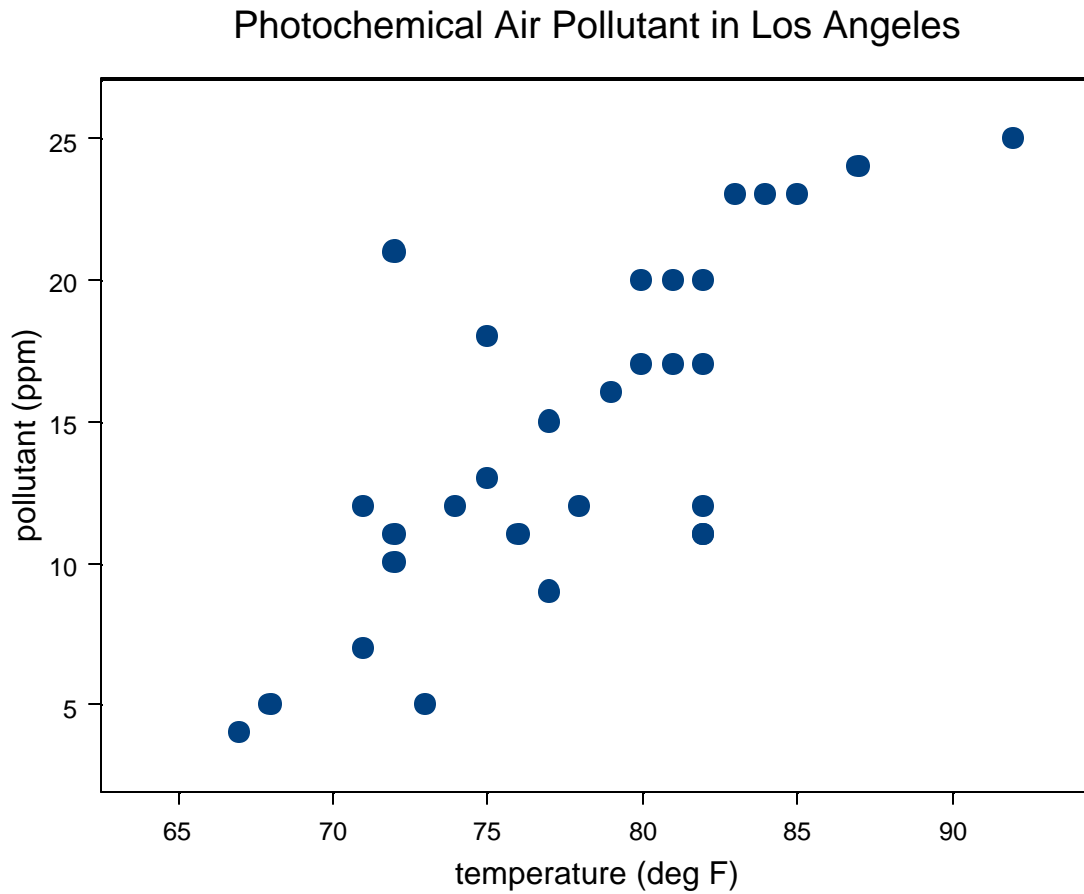
E.g. Side-by-side box plots.



2.3.6. Two-Way Scatter Plots

- For depicting the **relationship** between two numerical data types

E.g. 2-way scatter plot



2.3.7. Line Graphs

- Similar to scatter plots but only one y-value per x-value
- “Connect the dots”
- Often x-value is time