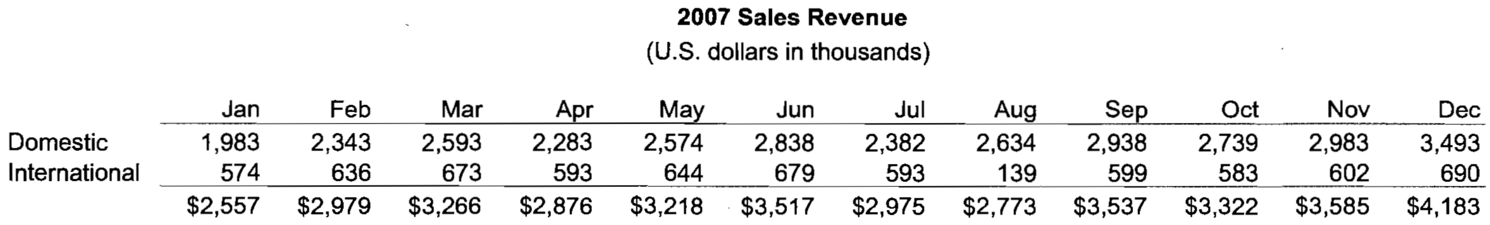
Introduction:

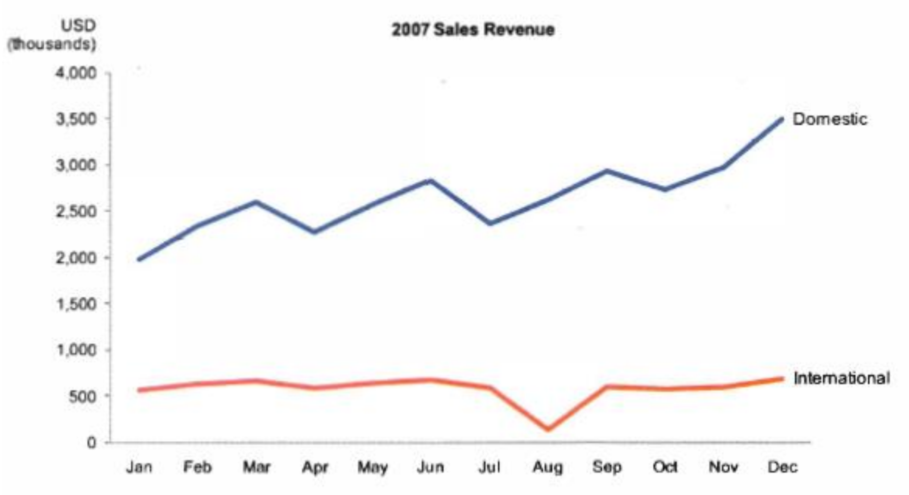
- Our brain is more accustomed to process visual stimuli and to recognize patterns in graphs than to find patterns in tables with columns of numbers.

  Experiment: Try to describe the data either using the table or the graph. Which takes more effort?



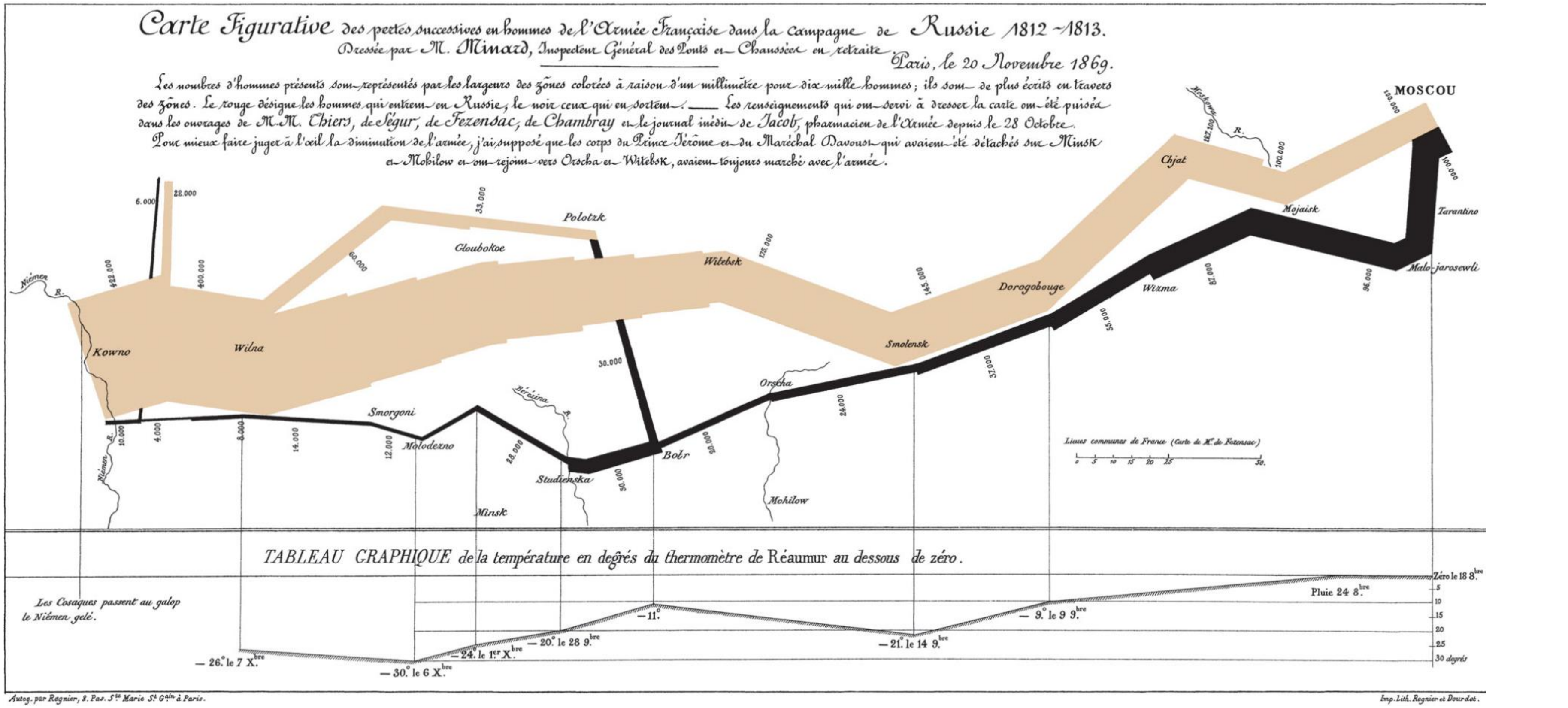
2007 Sales Revenue
(U.S. dollars in thousands)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Domestic	1,983	2,343	2,593	2,283	2,574	2,838	2,382	2,634	2,938	2,739	2,983	3,493
International	574	636	673	593	644	679	593	139	599	583	602	690
	$2,557	$2,979	$3,266	$2,876	$3,218	$3,517	$2,975	$2,773	$3,537	$3,322	$3,585	$4,183



Note: Common scale for both variables helps comparison.

- Minard’s depiction of the demise of Napoleon’s Grant Army in the Russian campaign June 23, 1812 to December 1812. An army of 422,000 left but only 10,000 returned.



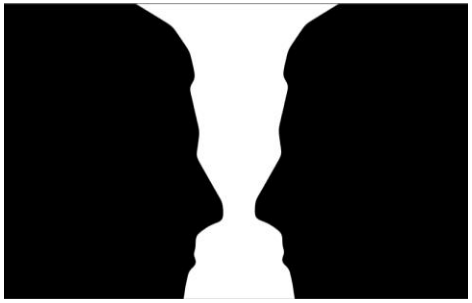
See also the R script **`MinardTroops.R`**. Note that the package **`HistData`** comprises of several historically interesting data sets.

- The Role of Graphics in Information Processing

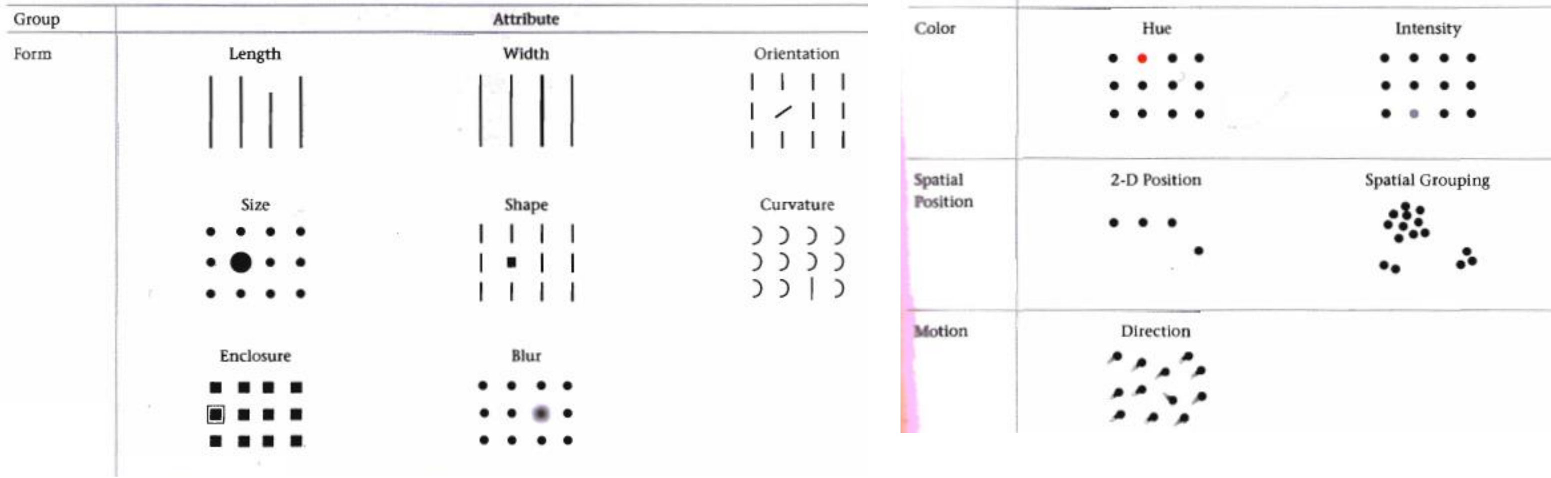
> ***Modern data graphics can do much more than simply substitute for small statistical tables. At their best, graphics are instruments for reasoning about quantitative information. Often the most effective way to describe, explore, and summarize a set of numbers – even a very large set – is to look at pictures of those numbers. Furthermore, of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful.***

From the Introduction to Edward R. Tufte, **1983**. *The Visual Display of Quantitative Information*. Graphics Press

- John Tukey quotes:
  - ***It is better of have a fuzzy answer to right questions than a precise answer to the wrong question.***
  - ***The nice thing about being a statistician*** *(or a geo-spatial data analyst)* ***is that I can play in everybody's backyard.***
- Note the changing focus of GISciences from spatial data handling to geo-spatial data analytics.
- Caution: Sometimes our visual perception of patterns may be tricked by illusions happening in our brains: Face or vase?

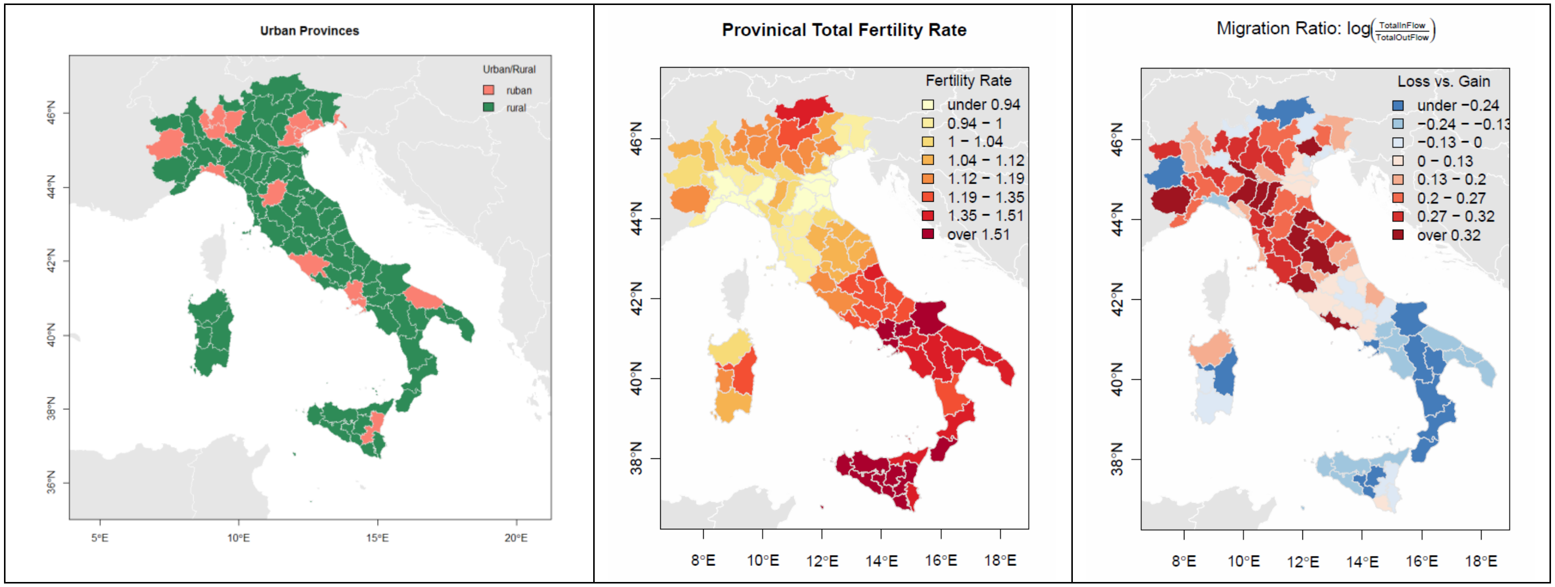


- Modern computer software supports the ***visual exploration of data*** and finding the most communicative way of presenting the ***information hidden*** within our data to gain meaningful knowledge about the underlying mechanisms that have generated our data.
- The science of semiology (study of how signs and symbols are processes) has developed a grammar of graphics. See Leland Wilkinson, **2005**. *The Grammar of Graphics*. 2nd edtion, Springer Verlag.
- The rules *The Grammar of Graphics* are implemented in R's library **`ggplot2`**.
- Elements of the grammar of graphics are:



- Color can be a powerful graphical tool (see http://colorbrewer2.org/) if your output medium permits its use.

  For choropleth map (coloring a set of regions) one distinguishes between [a] ***categorical*** map themes, [b] ***gradient*** continuous map themes, and [c] ***bi-polar*** of ***diverging*** map themes with a natural break point.

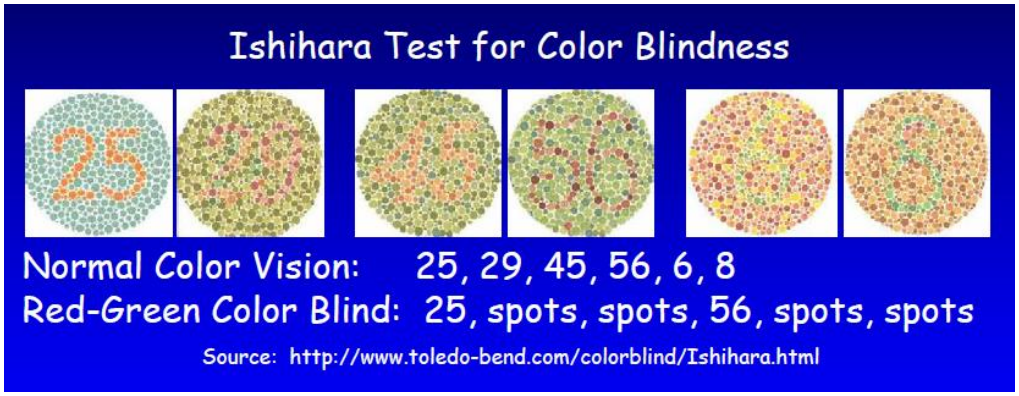

Urban Provinces

Provinical Total Fertility Rate

Migration Ratio: $\log\left(\frac{TotalInFlow}{TotalOutFlow}\right)$

See the R script **`MapTX.R`** for maps of the TX counties.

- However note: approximately 12 % of the population – predominately males – is colorblind (mainly unable to distinguish between red and green).


Ishihara Test for Color Blindness

Normal Color Vision: 25, 29, 45, 56, 6, 8

Red-Green Color Blind: 25, spots, spots, 56, spots, spots

Source: http://www.toledo-bend.com/colorblind/Ishihara.html

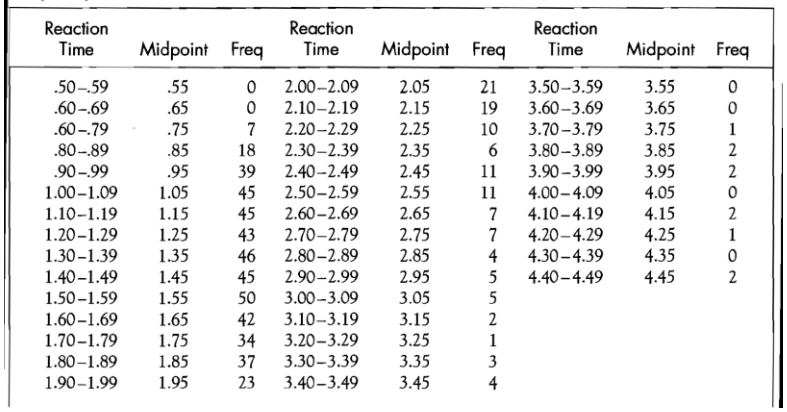
Frequency Distribution and Histogram

- Histograms are used to summarize the distribution of ***metrically scaled variables***. They display how frequently the observations in a data set will be within particular range of values.

- What does the distribution of variable tell us?
  - Reasonable ***value range***.
  - It gives an indication of the ***most frequent values***.
  - It gives an indication of the degree of ***spread*** and ***central location*** of the distribution.
  - It points to ***unusual observations*** with respect to the remaining sample observation.

- Since histograms are based on the aggregation of observations within given intervals ***information will be lost***. Original data values within an interval are substituted by the ***interval midpoint***.
- A frequency table is a numerical list underlying histograms.

Table 3.2
Frequency Distribution of Reaction in Times (in 10ths of seconds)



Reaction Time	Midpoint	Freq	Reaction Time	Midpoint	Freq	Reaction Time	Midpoint	Freq
.50–.59	.55	0	2.00–2.09	2.05	21	3.50–3.59	3.55	0
.60–.69	.65	0	2.10–2.19	2.15	19	3.60–3.69	3.65	0
.60–.79	.75	7	2.20–2.29	2.25	10	3.70–3.79	3.75	1
.80–.89	.85	18	2.30–2.39	2.35	6	3.80–3.89	3.85	2
.90–.99	.95	39	2.40–2.49	2.45	11	3.90–3.99	3.95	2
1.00–1.09	1.05	45	2.50–2.59	2.55	11	4.00–4.09	4.05	0
1.10–1.19	1.15	45	2.60–2.69	2.65	7	4.10–4.19	4.15	2
1.20–1.29	1.25	43	2.70–2.79	2.75	7	4.20–4.29	4.25	1
1.30–1.39	1.35	46	2.80–2.89	2.85	4	4.30–4.39	4.35	0
1.40–1.49	1.45	45	2.90–2.99	2.95	5	4.40–4.49	4.45	2
1.50–1.59	1.55	50	3.00–3.09	3.05	5			
1.60–1.69	1.65	42	3.10–3.19	3.15	2			
1.70–1.79	1.75	34	3.20–3.29	3.25	1			
1.80–1.89	1.85	37	3.30–3.39	3.35	3			
1.90–1.99	1.95	23	3.40–3.49	3.45	4			

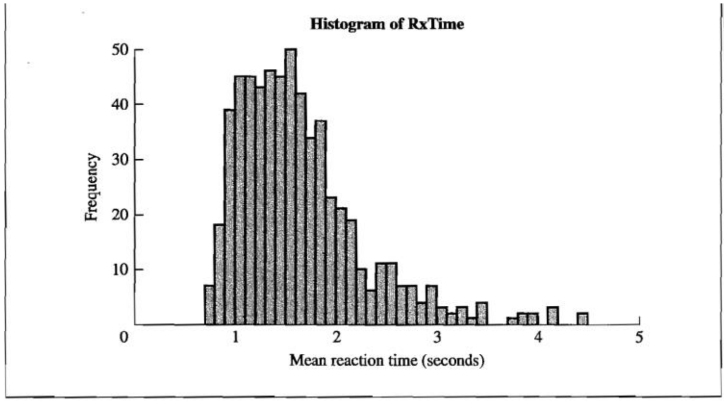


Figure 3.1
Plot of reaction times against frequency

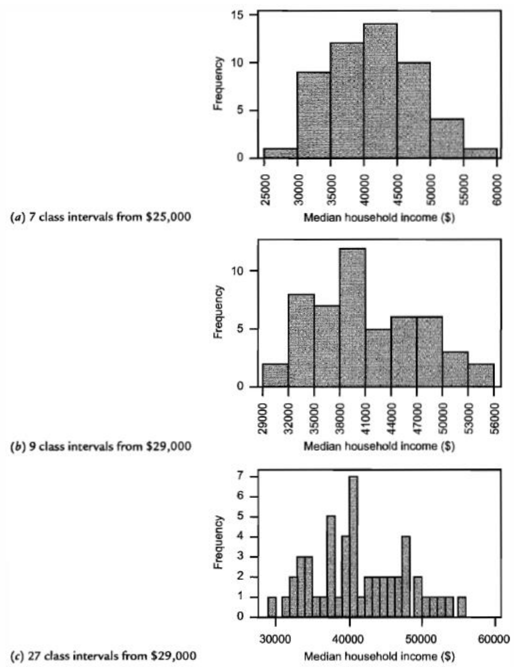
- Construction guidelines for class intervals:
    - The intervals cover the ***whole support*** of the data, they do not overlap and they do not exhibit gaps.
    - The intervals are on the x-axis and the frequencies (either counts or relative frequencies) are on the y-axis.
    - Select the appropriate number of interval classes. Some suggestions for the number of classes *k* in the literature are
        - $k = 1 + 3.3 \cdot \log_{10} n \;\; \text{for } n < 200$
        - The R function **`hist()`** gives three alternative options for the interval width h

Sturges' rule: $h = \text{range}(X) \big/ \left(\log_2 n + 1\right)$

Scott's rule: $h = 3.5 \cdot s \cdot n^{-1/3}$ with s being the estimated standard deviation

Freedman's rule: $h = 2 \cdot IQR \cdot n^{-1/3}$ with IQR being the interquartile range

- Selection of inappropriate number of classes
  - If the number of classes is ***too small*** it can mask critical characteristics of an underlying distribution.
  - If the number of classes is ***too large*** typical data clusters become diluted.
- Respect ***natural breaks***, such as *zero* degrees Celsius or a *ph*-level of 7.
- In general, the interval width needs to be ***constant***.
  This allows proper comparison of the frequencies among classes.
  Don't make the error shown below:



(*a*) 7 class intervals from $25,000

(*b*) 9 class intervals from $29,000

(*c*) 27 class intervals from $29,000

FIGURE 2-12. Median household income in U.S. states.

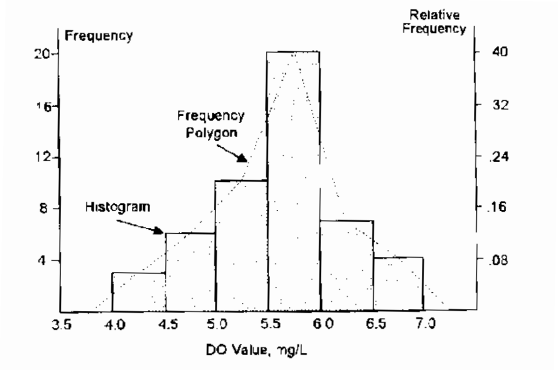


FIGURE 2-1. Histogram and frequency polygon.

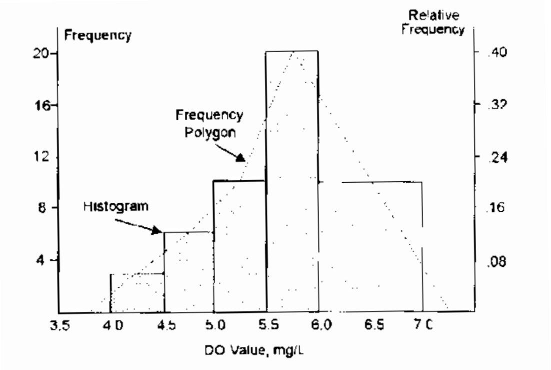


FIGURE 2-2. Histogram with unequal intervals

- In general, use the smallest and largest of the observation as lower and upper bound. If more than one histogram of the same data (e.g., city A and city B) is plotted side-by-side use a ***common*** classification scale.

  If there are a few far outlining observations it is advisable to plot first all data simultaneously and then ***zoom in a second graph*** into the bulk of the observations to gain a better ***visual resolution***.

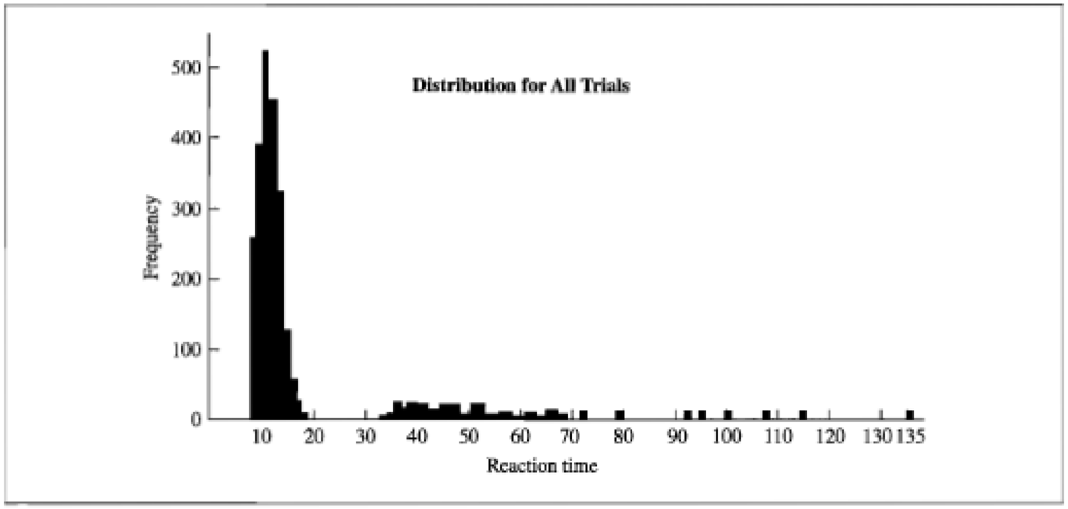
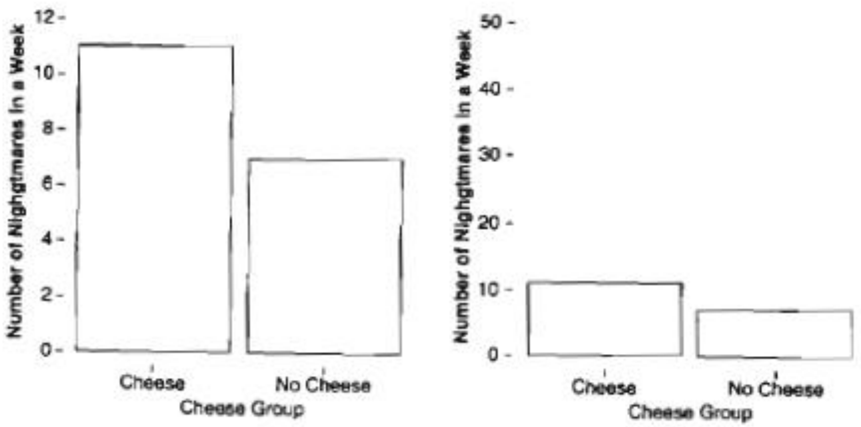


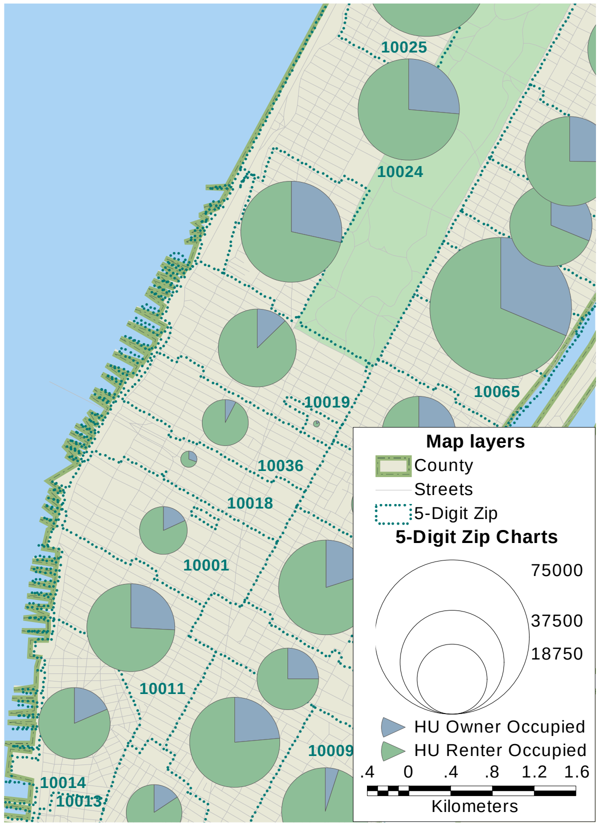
Figure 3.10
Frequency distribution of Bradley's reaction time data

Bar/Pie Charts

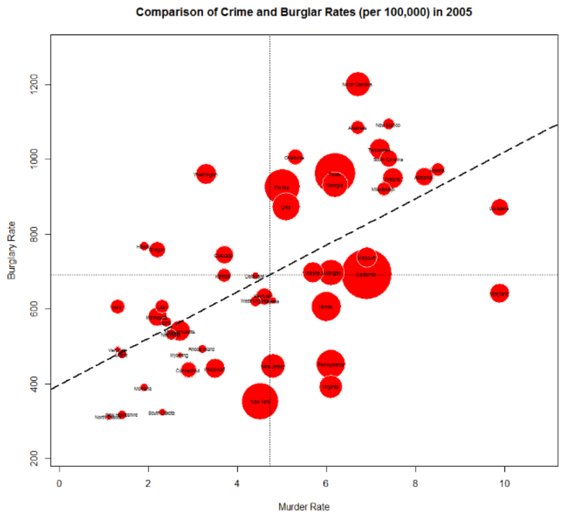
- Bar charts plot the frequency or other summary statistics (y-axis) against the representation of categorical variables (x-axis).

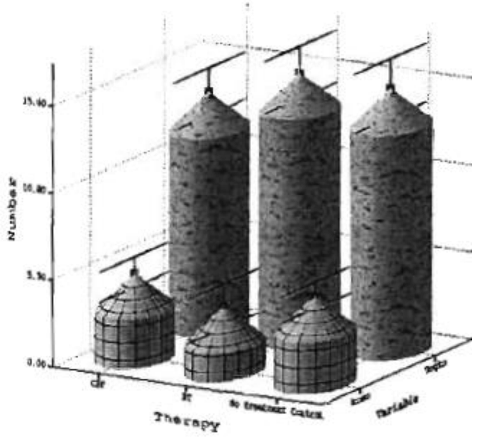


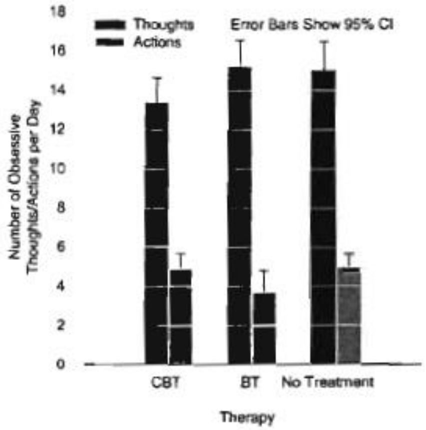
- Warning: Make sure to properly scale the y-axis to highlight differences in the counts or summary statistic.
- Since the order of categories is irrelevant it is good practice to first display the ***most frequent*** category first, then the next frequent category and to forth.
- I am ***not a big fan*** of ***pie charts***. They do not communicate the information in a professional manner. However, sometimes for efficient comparison purposes, they are quite useful. In particular, since their size can also be varied in a meaningful way.



- For bubble plots the circle radius r needs to be proportional the relative sizes variable: $r = \sqrt{sizeVar/\pi}$. See the R-script **BUBBLEPLOT.R** from the website http://www.flowingdata.com.
- If information can be displayed 2-dimensional. 3d charts frequently distract form the underlying information in a graph:

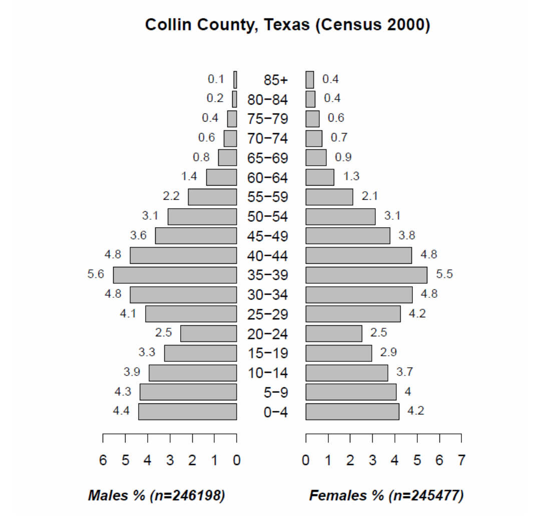




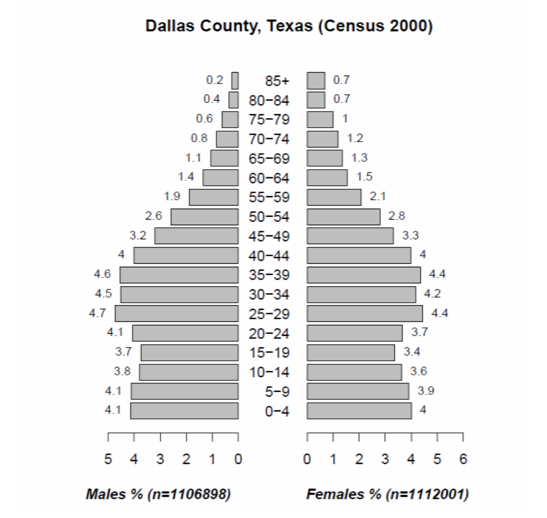


Side-by-Side Graphs:

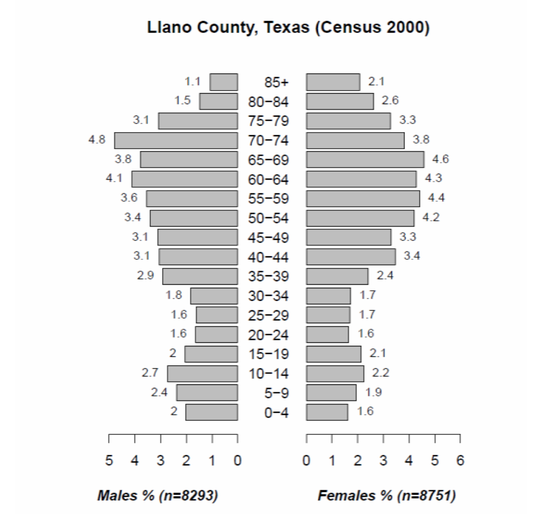
- Side-by-side graphs break observations apart by a categorical variable. They are histograms rotated by 90 degrees. Example: Selected Texas counties' population pyramids



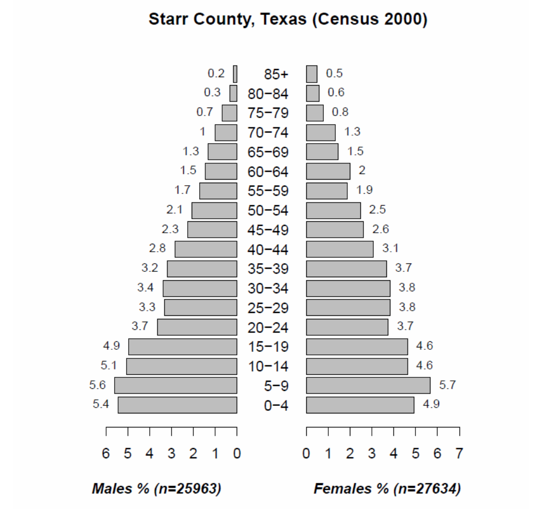
empty nesters



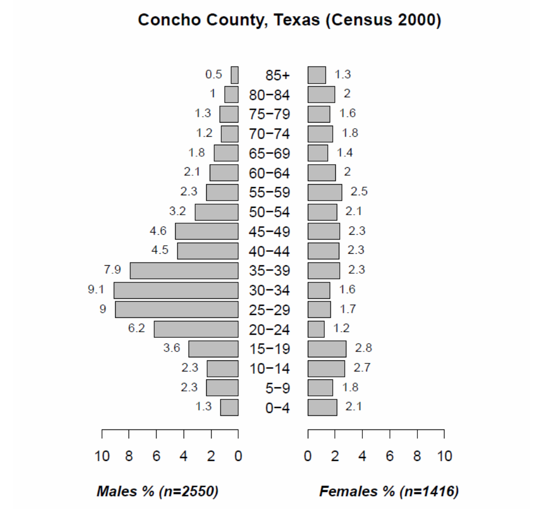
balanced ?



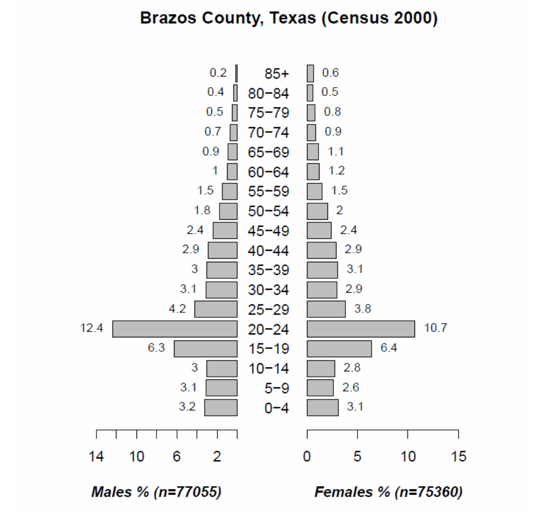
retirees



TX-Mexico border Hispanics



male detention center

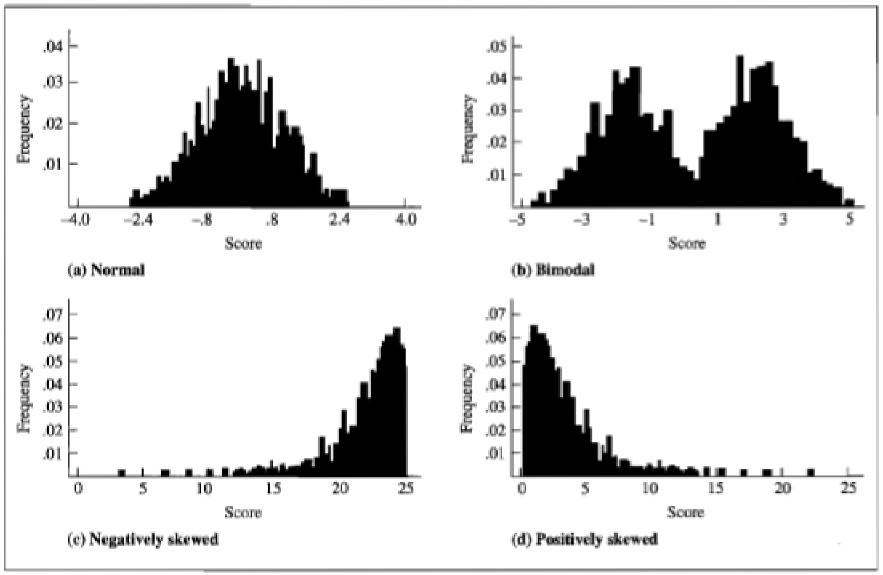


Texas A&M University

Note: ***common scale*** is used for males and females and ***relative frequencies*** are used.

The Shape of Distributions

- Distributions can be distinguished with regards the ***balance*** of their left and right tails:
  - ***Symmetric*** distributions. Tails are balanced into either direction from a central value.
  - ***Negatively*** skewed distributions (long tail into the direction small values)
  - ***Positively*** skewed distributions (long tail into the direction of large values).
    These distributions frequently emerge for variables with a binding lower origin (like zero income).
  - Extreme skewness may hint at ***outliers*** that do not match the rest of the observed data.
    ⇒ A different data generating process may have generated the different observations.
- The number of meaningful clusters of observations is described by the term modality:
  - Uni-modality refers to just one peak
  - Bi-modality refers to two outstanding peaks
  - Multimodality refers to more than two outstanding peaks.
  - Skewness refers to a longer tail in one direction and a concentrated distribution in the other direction.



(a) Normal

(b) Bimodal

(c) Negatively skewed

(d) Positively skewed

Figure 3.9
Shapes of frequency distributions: (a) Normal; (b) Bimodal; (c) Negatively skewed; (d) Positively skewed

Cumulative Frequencies and Ogive

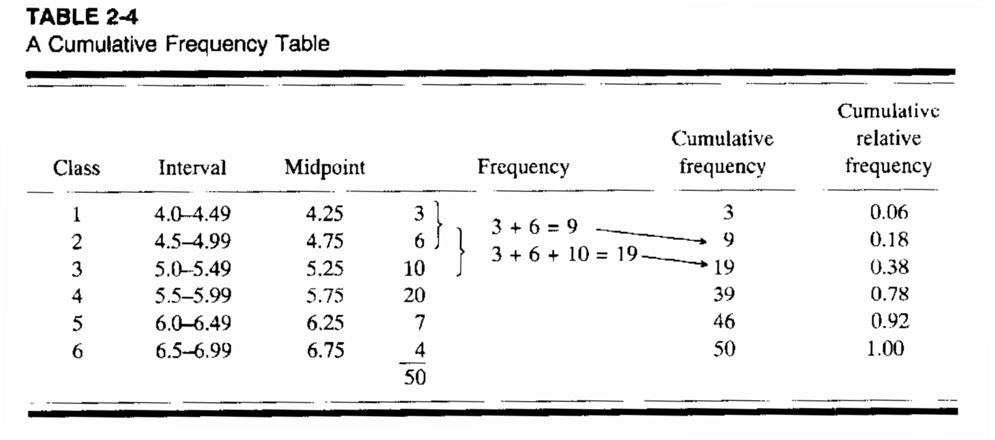


TABLE 2-4
A Cumulative Frequency Table

Class	Interval	Midpoint	Frequency		Cumulative frequency	Cumulative relative frequency
1	4.0–4.49	4.25	3		3	0.06
2	4.5–4.99	4.75	6	3 + 6 = 9 →	9	0.18
3	5.0–5.49	5.25	10	3 + 6 + 10 = 19 →	19	0.38
4	5.5–5.99	5.75	20		39	0.78
5	6.0–6.49	6.25	7		46	0.92
6	6.5–6.99	6.75	4		50	1.00
			50			

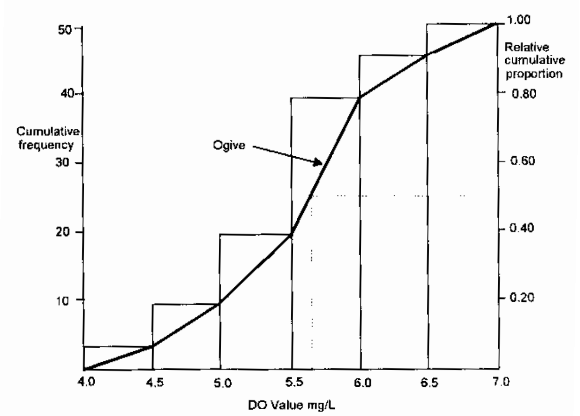


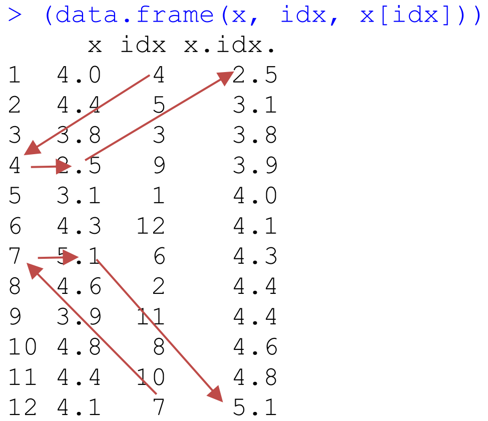
FIGURE 2-7. Cumulative frequency histogram and ogive.

- The ***ogive*** allows estimating the percentage of observations below a threshold value. E.g., 50% of the lakes have a DO value below 5.65 mg/L.
- What happens as the number of sample observations increases?

  ⇒ Since we can make the class intervals (bins) increasingly finer the *ogive* will approach the underlying cumulative population distribution.
- A problem with grouped (i.e., aggregated) data is that we lose information. We do not know the exact values within each group, the best we can do is to represent each grouped observation by its group mean.

Quantiles and Percentiles

- Technically, quantiles and percentiles are generated from a ***sorted list*** of the original data points $x_{[1]} \leq x_{[2]} \leq x_{[3]} \leq \cdots \leq x_{[n-1]} \leq x_{[n]}$ where each observations has an assigned rank $i \in \{1,2,\ldots,n\}$, with $i=1$ for the smallest observation and $i=n$ for the largest observation.
- Example of **`order( )`** function in R-script **`QuantileDemo.r`**:

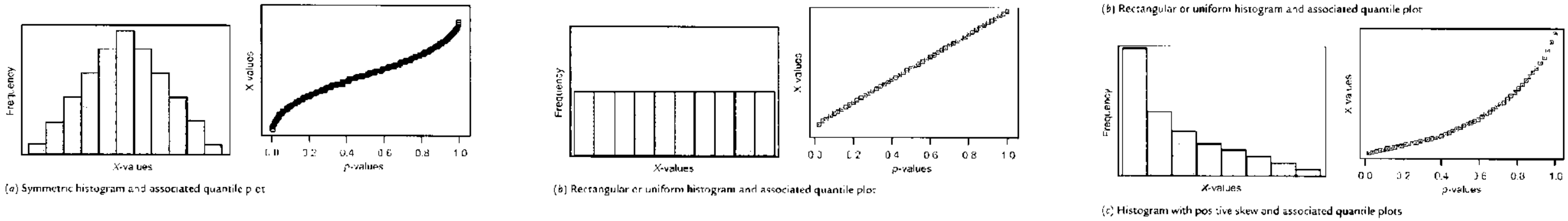


```
> (data.frame(x, idx, x[idx]))
      x idx x.idx.
1   4.0   4    2.5
2   4.4   5    3.1
3   3.8   3    3.8
4   2.5   9    3.9
5   3.1   1    4.0
6   4.3  12    4.1
7   5.1   6    4.3
8   4.6   2    4.4
9   3.9  11    4.4
10  4.8   8    4.6
11  4.4  10    4.8
12  4.1   7    5.1
```

- For a give data value $\boldsymbol{x_{[i]}}$ the ***percentile*** measures the proportion of sample observations less or equal to $\boldsymbol{x_{[i]}}$. The rank $\boldsymbol{i}$ and the number of observations $\boldsymbol{n}$ are used for their calculation.
- In its general form a ***percentile*** is calculated as $p\left(x_{[i]}\right) = \dfrac{i-\alpha}{n+(1-\alpha)-\alpha}$ where $x_{[i]}$ is the i-th sorted data value with $\mathbf{0 \leq \boldsymbol{\alpha} \leq 1}$
  - A value of $\alpha = 0.0$ gives $p\left(x_{[i]}\right) = \dfrac{i}{n+1}$ with $p \in \left[\frac{1}{n+1}, \frac{n}{n+1}\right]$.

- A value of $\alpha = 0.5$ gives $p\left(x_{[i]}\right) = \dfrac{i-0.5}{n}$ with $p \in \left[\frac{0.5}{n}, \frac{n-0.5}{n}\right]$.
- A value of $\alpha = 1.0$ gives $p\left(x_{[i]}\right) = \dfrac{i-1}{n-1}$ with $p \in [0,1]$.

- This adjustment coefficient α links the percentile back to an underlying hypothetical distribution:
  Justification for $\boldsymbol{\alpha < 1}$: if additional sample observations become available, then it would be possible to observe data values that are smaller or larger than the most extreme observations $x_{[1]}$ and $x_{[n]}$ in the current sample.
- The R function `ppoints(n,a)` calculates the quantiles. Try the function for $\alpha \in \{0, 0.5, 1\}$ to understand its behavior.
- A ***quantile*** is that data value $\boldsymbol{x_{[i]}}$ of a distribution, which is associated with a particular percentile point.
  - For instance, if $\alpha = 1$ has been selected for the percentiles, then $x_{[i]} = q\left(p\left(x_{[i]}\right)\right)$.
  - For any other percentile not obtained in the dataset (a gap value), the quantile is linearly interpolated by:
    $q(p) = \gamma \cdot x_{[i]} + (1-\gamma) \cdot x_{[i+1]}$ with $p\left(x_{[i]}\right) \le p \le p\left(x_{[i+1]}\right)$ and $0 \le \gamma \le 1$
- Specific distributions (normal, uniform, skewed) have characteristic quantile plots:



(*a*) Symmetric histogram and associated quantile plot

(*b*) Rectangular or uniform histogram and associated quantile plot

(*b*) Rectangular or uniform histogram and associated quantile plot

(*c*) Histogram with positive skew and associated quantile plots

- Important quantiles are:
  - 0.25 quantile also called Q_1 quartile (25 percent of the observations are smaller or equal to this quantile value)
  - 0.50 quantile also called the median (50 of the observations are smaller or larger than the given quantile value)
  - 0.75 quantile also called Q_3 quartile (75 percent of the observations are smaller or equal to this quantile value and 25 percent of the observations are larger than this value)
  - A measure of spread is the inter-quartile range: $IQR = Q_3 - Q_1$

Box-Plots

- Construction of the box-plot
  - Draw a box from Q_1 to Q_3. Mark the median $\boldsymbol{Q_2}$ in the center of the box with a line.

- Definition of adjacent values $x_{low}^{adj} = \min\left(x_{[i]} \in \left(Q_1, Q_1 - 1.5 \cdot IQR \right) \text{ and } x_{[i]} \text{ in dataset} \right)$ and $x_{high}^{adj} = \max\left(x_{[i]} \in \left(Q_3, Q_3 + 1.5 \cdot IQR \right) \text{ and } x_{[i]} \text{ in dataset} \right)$.

  The term $x \in (a,b)$ means, all x-values in the interval between a and b.

  Draw the "fences" so they just include the smallest and largest data values x_{low}^{adj} and x_{high}^{adj}, respectively.

- Outliers are in the interval $\left[1.5 \cdot IQR, 3.0 \cdot IQR\right]$ starting from Q_1 below or Q_3 above, respectively.

  Severe outliers are beyond that range $(> 3.0 \cdot IQR)$

- Comparison of the Histogram with the Box-Plot:

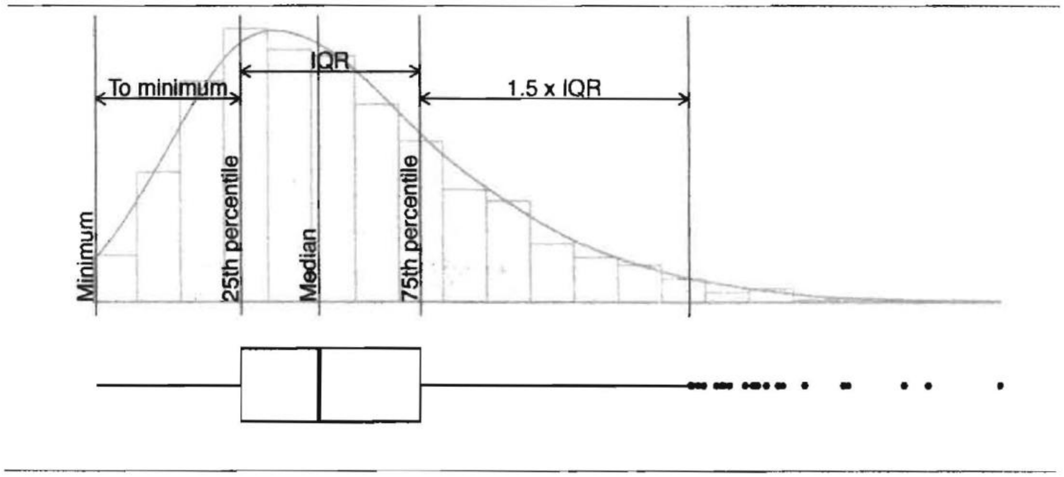
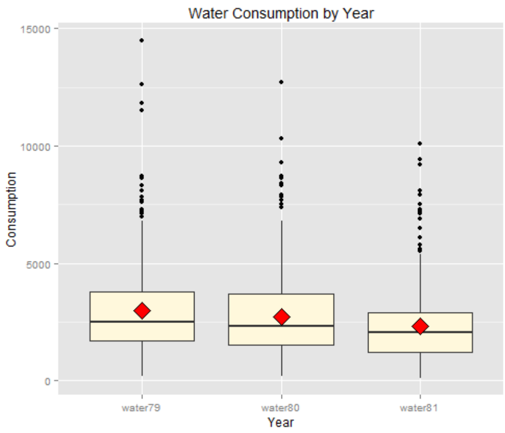


Figure 6-16. Box plot compared to histogram and density curve

  - The box-plot is not able to show multimodal distributions.
  - The box-plot highlights the tails of a distribution better (in particular for skewed distributions and distributions with outliers).

- Use of box-plots:
  - Easy visual description of the distribution of a variable and potential outliers
  - Comparison of distributions for several variables side-by-side.

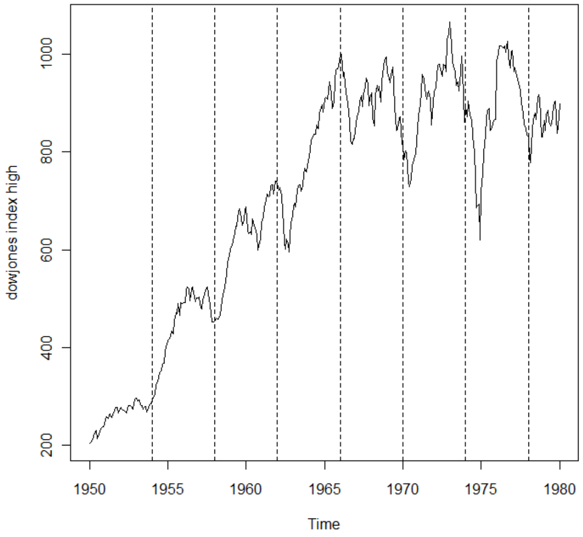


Time Series Plots/Line Graphs (skipped)

- Def. Time Series: A time series is a sequence of observations $Y_1, Y_2, \ldots, Y_t, \ldots, Y_T$ that are collected over successive increments of time.
- Interpretations:
  - Observations in a time series have the natural order.

  - The subscript *t* denotes the time when the observation was collected, thus a smaller index indicate an earlier observation.

- Usually the time interval Δt between observations is constant.
- Observations in time series usually have a periodicity, e.g., seasonally for months in a year or diurnal for hours in a day.
- Other examples: Waypoints along a route can be represented as a time series. A GPS in tracking mode takes measurements at fix intervals.
- It is important for time-series that the measurement rules remain constant throughout the series. Example: Change of the definition of inflation over time.
- Small gaps in a time-series due to missing values can be estimated by interpolating between values due to the inherent persistency (serial autocorrelation) in a series or by typical values borrowed from other periods.



Important Rfunctions for Plotting:

- See **MaindonaldUsingR.pdf** chapter 4.
- `barplot(x)` for bar plots
- `pie(x)` for pie charts

- **`boxplot(x)`** for boxplots.
- **`hist(x)`** for histograms
- **`stem(x)`** for stem-and-leaf plots
- **`ppoints(x,a)`** calculates a sequence of p-values
- **`table(x)`** ungrouped (and grouped) frequency tables
- **`stripchart(x,method="stack")`** for dot plots
- **`quantile(x)`** for quantiles and **`ppoints(n,a)`**

Design Guidelines

- Be clear what information goes on the x-axis (ordinate) and what on the y-axis (abscissa).
- Be clear that histograms, stem-and-leaf and trend graphs are for metric variables and bar charts are for categorical variables.
- Make sure that your graphs communicate its intended message utilitarian, neat and orderly (this usually translates into a minimal amount of applied ink).
- Supply a title for each graph and a longer caption underneath it.
- Always label the axis properly.
- Don't add nonessential material.
- Avoid 3d, pie charts and other “eye candy” (search Google for “bad graphs” for a good laugh)