CS 495:
Data Visualization for Data Scientists

Waleed A. Yousef, Ph.D.,

Human Computer Interaction Lab.,
Computer Science Department,
Faculty of Computers and Information,
Helwan University,
Egypt.

March 24, 2019

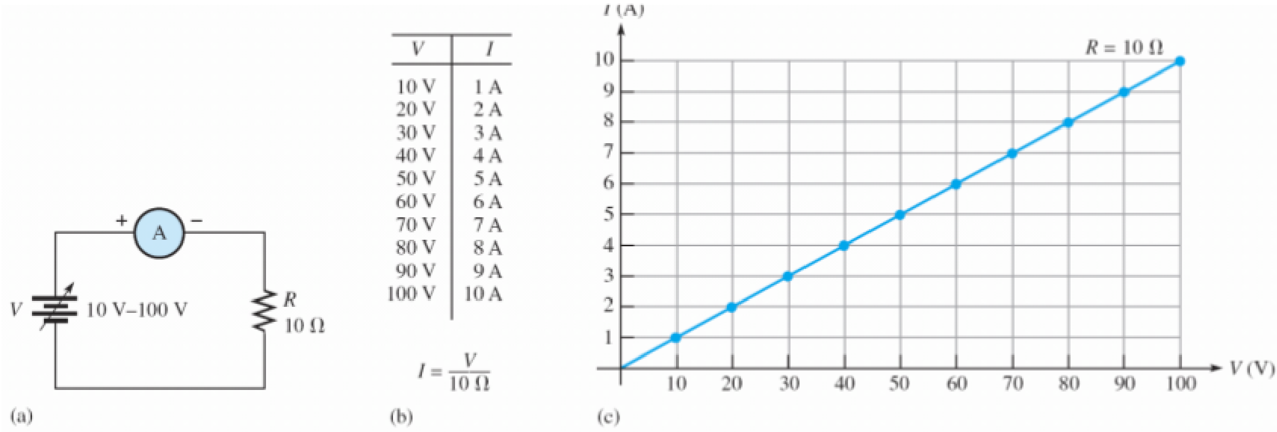
Prologue and Motivation

- "A picture is worth a thousand words" (English idiom).
- Recent research suggests that:
  "*Retina communicates to brain at 10 million bits per second. 40 words per second are read at 10 sec.; call it 1000 bits/sec. which is 1/10,000*"[1]

[1] https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0002NC

Data Visualization for Exploration

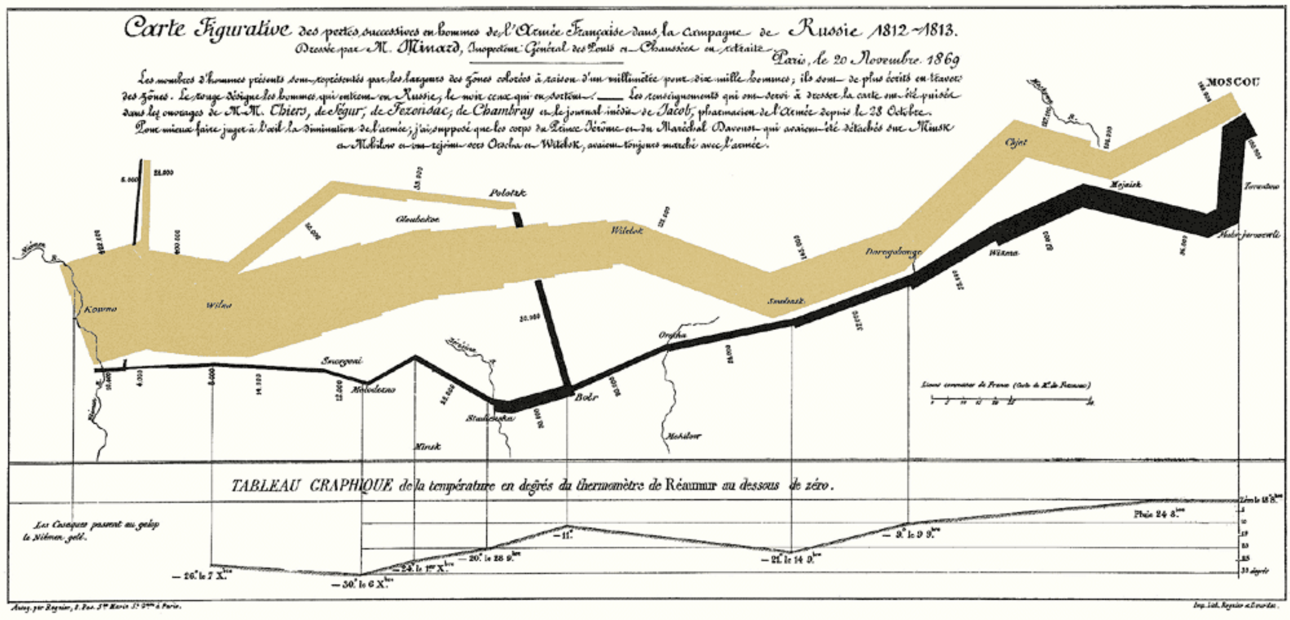
The discovery of the very classic Ohm's law



- Then, comes Statistics, Statistical Learning, Pattern Recognition, to formalize the observed relationship: model, regression, $p-$values, variance, confidence interval, etc.

Data Visualization for Illustration and Presentation

Invasion/retreat of French army to/from Russia:

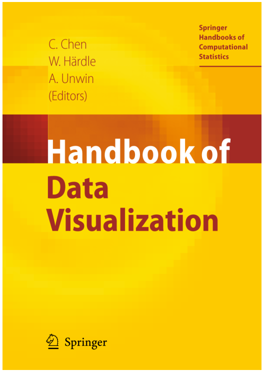


“Vivid historical content and brilliant design combine to make this one of the best statistical graphics ever” (Tufte, 2006, P. 122)

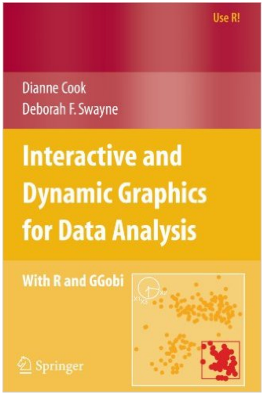
Course Objectives

- Data Visualization:
  - for exploring, extracting secrets, and understanding
    - build intuition and insight.
    - allow you getting the feeling of the patterns, secrets, hiding in data.
    - understand your data before any mathematical treatment.
  - for illustration, displaying, and conveying what has been explored.
- Linking to real life problems.
- Coding and scientific computing.
- We will emphasize on the foundations than the evolving technology.
- This course is just a very interesting voyage in high dimensions and hyperspace. Please, prepare your baggage, video cam, juice, and say cheese.

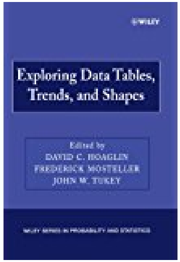
Texts, References, and Prerequisites



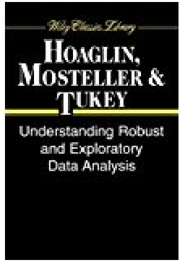
Chen, C.-h., Härdle, W., Unwin, A., 2008. Handbook of data visualization. Springer, Berlin



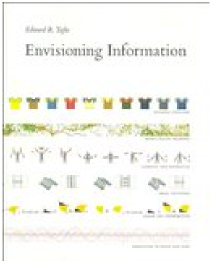
Cook, D., Buja, A., Lang, D. T., Swayne, D. F., Hofmann, H., Wickham, H., Lawrence, M., 2007. Interactive and Dynamic Graphics for Data Analysis: With R and GGobi. Springer Science & Business Media



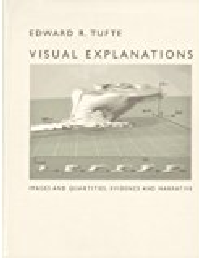
Hoaglin, D. C., Mosteller, F., Tukey, J. W., 2000. Understanding robust and exploratory data analysis, wiley clas Edition. Wiley, New York



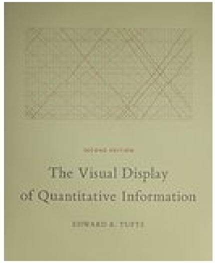
Hoaglin, D. C., Mosteller, F., Tukey, J. W., 1985. Exploring data tables, trends, and shapes. Wiley, New York



Tufte, E. R., 1990. Envisioning Information. Graphics Press, Cheshire, Conn

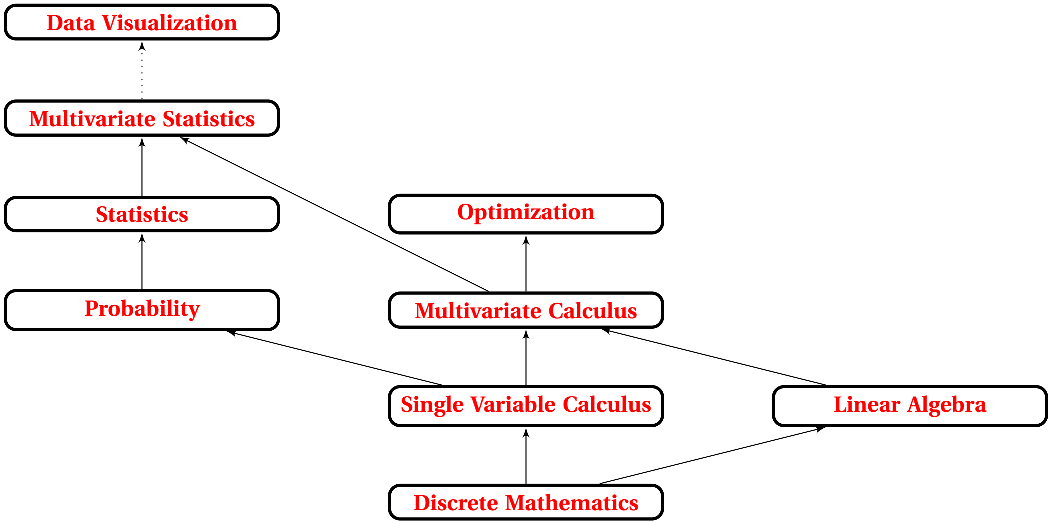


Tufte, E. R., 1997. Visual explanations : images and quantities, evidence and narrative. Graphics Press, Cheshire, Conn



Tufte, E. R., 2001. The visual display of quantitative information, 2nd Edition. Graphics Press, Cheshire, Conn

Tufte, E. R., 2006. Beautiful evidence. Graphics Press, Cheshire, Conn


Data Visualization
Multivariate Statistics
Statistics
Optimization
Probability
Multivariate Calculus
Single Variable Calculus
Linear Algebra
Discrete Mathematics

Contents

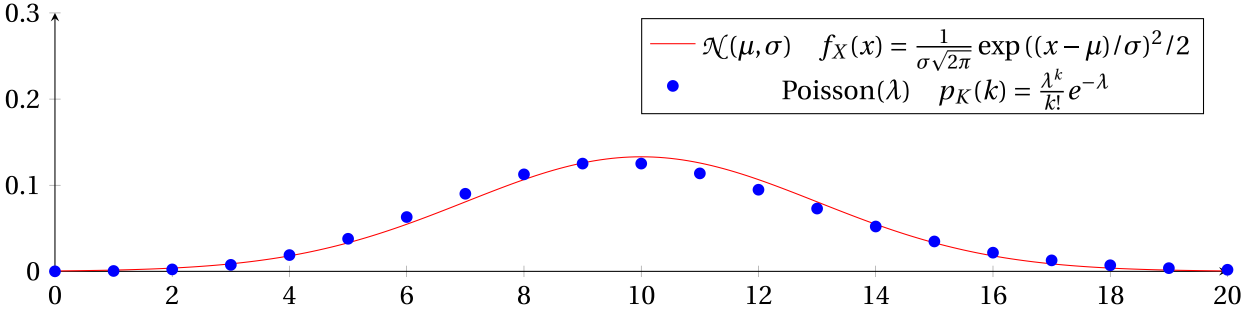
Part I

Exploration

Chapter 1

Some Necessary Probability and Statistics

1.1 Samples from Discrete and Continuous Distributions



- Here, $\mu = 10$, $\sigma = 3, \lambda = 10$ (how do you know from figure?)
- $P(X = x) = 0, P(K = k) \neq 0$.
- How samples look like?
- What about cluttering (observations overlaying each other).

1.2 Cumulative Distribution Function (CDF)

$$
\begin{aligned}
F(x) &= P(X \leq x) \\
&= \int_{-\infty}^{x} f(u)\ du = P(X < x) && \text{(cont. var.)}
\end{aligned}
$$

Definition 1 **(F^{-1})** *: The p^{th} quantile is defined as, the value x_p of the r.v. that satisfies $F\left(x_p\right) = p$.*

- If F is monotonically (strictly) increasing, the pth quantile is unique (see figure).
- $F^{-1}(.5)$ is the median.
- $F^{-1}(.25)$ and $F^{-1}(.75)$is the lower and upper quartile.

Example 2 *Suppose*

$$F(x) = x^2,\ 0 \leq x \leq 1,$$

$$
\begin{aligned}
x_p^2 &= p, \\
x_p &= \sqrt{p}, \\
x_{.5} &= \sqrt{.5} = .707 \\
x_{.25} &= \sqrt{.25} = .5 \\
x_{.75} &= \sqrt{.75} = .866
\end{aligned}
$$

1.3 Normal Distribution

Corollary 3 *If $X \sim \mathcal{N}(\mu, \sigma)$ and $Z = \sim \mathcal{N}(0,1)$ (a standard normal), then*

$$P(Z < z) = \int_{-\infty}^{z} f_Z(u)\, du = \Phi(z)$$

$$\Phi(z) = 1 - \Phi(-z)$$

$$\frac{(X-\mu)}{\sigma} \sim Z$$

$$P(X < x) = P\left(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}\right) = P\left(Z < \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Example 4 *[σ and μ]:*

$$\begin{aligned}
P\left(\left|X-\mu\right| < \sigma\right) &= P\left(-\sigma < X-\mu < \sigma\right) \\
&= P\left(-1 < \frac{X-\mu}{\sigma} < 1\right) \\
&= P\left(-1 < Z < 1\right) \\
&= \Phi\left(1\right) - \Phi\left(-1\right) \\
&= .68 \\
P\left(\left|X-\mu\right| < 2\sigma\right) &= \Phi\left(2\right) - \Phi\left(-2\right) \\
&= .9545, \\
P\left(\left|X-\mu\right| < 3\sigma\right) &= \Phi\left(3\right) - \Phi\left(-3\right) \\
&= .9973 \qquad \text{(almost all the probability measure)}
\end{aligned}$$

1.4 Quantile Estimation, Outliers Cutoff, and Thick Tails

The ordered statistic $x_{(p=i.d)}$ is defined by interpolation as:

$$
\begin{aligned}
x_{(i.d)} &= x_{(i)} + d(x_{(i+1)} - x_{(i)}) = (1-d)x_{(i)} + d\,x_{(i+1)} && = \widehat{F}^{-1}(p) && (p^{\text{th}} \text{ quantile})\\
x_{((n+1)/2)} &= \begin{cases} x_{((n+1)/2)}, & n \text{ is odd.}\\ x_{(n/2+1/2)} = (1/2)(x_{(n/2)} + x_{(n/2+1)}) & n \text{ is even.}\end{cases} && = \widehat{F}^{-1}(0.5) && (\text{median: M})\\
x_{((1+(n+1)/2)/2)} &= x_{((n+3)/4)} && = \widehat{F}^{-1}(0.25) && (\text{lower quartile: } Q_L)\\
x_{(((n+1)/2+n)/2)} &= x_{((3n+1)/4)} && = \widehat{F}^{-1}(0.75) && (\text{upper quartile: } Q_U)\\
W_L &= \min x_i \geq Q_L - 1.5(Q_U - Q_L) && && (\text{sample lower cutoff})\\
W_U &= \max x_i \leq Q_U - 1.5(Q_U - Q_L) && && (\text{sample upper cutoff})
\end{aligned}
$$

Example 5

34	*35*	*36*	*37*	*45*	*52*	*56*	*58*	*66*	*68*	*74*	*90*	*100*	*145*
1	2	3	4	5	6	7	8	9	10	11	12	13	14

Rank of M, QL, QU is 7.5, 4.25, 10.75

$M = 56 + 0.5(58 - 56) = 57$

$Q_L = 37 + 0.25(45 - 37) = 39$

$Q_U = 68 + 0.75(74 - 68) = 72.5$

$d_Q = (72.5 - 39) = 33.5$

$Q_L - 1.5\ d_Q = 39 - 1.5 \times 33.5 = -11.25 \qquad W_L = 34$

$Q_U + 1.5\ d_Q = 72.5 + 1.5 \times 33.5 = 122.75 \qquad W_U = 100$

Example 6 (meaning of quantile from $X \sim \mathcal{N}(\mu, \sigma)$) :

$$p = F(x_p) = P(X < x_p) = \Phi\left(\frac{x_p - \mu}{\sigma}\right)$$

$$F^{-1}(p) = x_p = \mu + \left(\Phi^{-1}(p)\right)\sigma$$

$$Q_L = F^{-1}(0.25) = \mu - 0.6745\sigma$$

$$Q_U = F^{-1}(0.75) = \mu + 0.6745\sigma$$

$$d_Q = 1.349\sigma$$

$$W_L = Q_L - 1.5d_Q = \mu - 2.698\sigma$$

$$W_U = Q_U + 1.5d_Q = \mu + 2.698\sigma$$

$$P(X < W_L) + P(X > W_U) = 2P(X < W_L) = 2\Phi\left(\frac{(\mu - 2.698\sigma) - \mu}{\sigma}\right) = 2\Phi(-2.698) = 0.00698$$

So, a sample (patch) of 1000 obs. will have almost 7 obs. outside the cutoffs.

Definition 7 (Hoaglin et al. (2000)) :

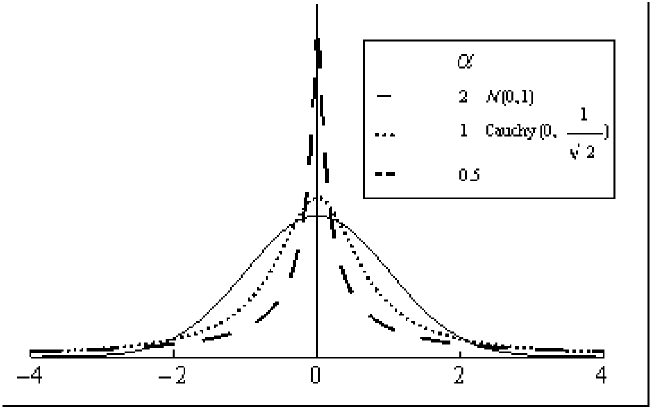
Outlier *observation with different underlying behavior as compared with the bulk of the data which deserves more investigation. The cutoffs W_L and W_U will be arbitrarily used for outlier detection. Outliers could be:*

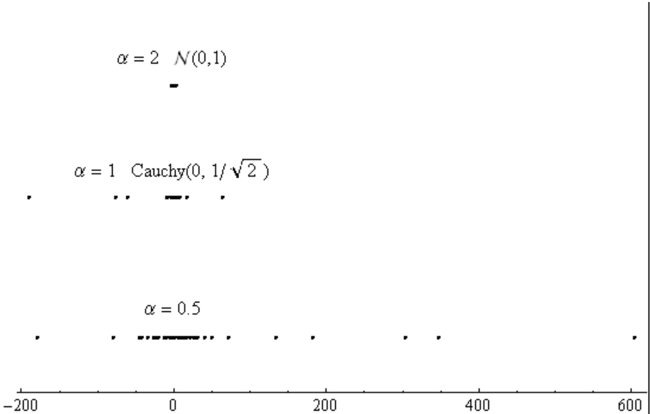
- *false value due to measurement error.*
- *right value due to thick tail.*

Resistance *insensitivity to misbehavior in data. A resistant method produces results that change only slightly when small part of the data is replaced by new numbers, possibly very different from the original ones.*

Robustness *insensitivity to departure from assumptions surrounding an underlying probabilistic model.*

Example 8 (Stable Distributions: thick tailed) :





1.5 Transformation and Log-scale

Chapter 2

History and Introduction

2.1 Evolution of Data Visualization

2.2 Types of Variable

Quantitative, where some measure is given as a value; e.g.,$X = 1, 3, -2.5$.

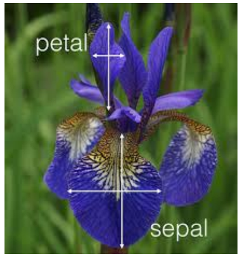
Qualitative (or Categorical), where no measures or metrics are associated; e.g., $X = Diseased, Nondiseased$.

Ordered Categorical; e.g., $X = small, medium, ...$. The variable $X \in \mathcal{G}$, a set of possible values.

Example 9 (`iris` dataset) *: (150 observations, by R. A. Fisher, the father of Statistics)*

Index	*SepalLength*	*SepalWidth*	*PetalLength*	*PetalWidth*	*Class*
1	5.1	3.5	1.4	0.2	*Iris-setosa*
2	4.9	3	1.4	0.2	*Iris-setosa*
3	4.7	3.2	1.3	0.2	*Iris-setosa*
4	4.6	3.1	1.5	0.2	*Iris-setosa*
5	5	3.6	1.4	0.2	*Iris-setosa*
6	5.4	3.9	1.7	0.4	*Iris-setosa*
7	4.6	3.4	1.4	0.3	*Iris-setosa*
8	5	3.4	1.5	0.2	*Iris-setosa*
9	4.4	2.9	1.4	0.2	*Iris-setosa*
10	4.9	3.1	1.5	0.1	*Iris-setosa*
⋮					
51	7	3.2	4.7	1.4	*Iris-versicolor*
52	6.4	3.2	4.5	1.5	*Iris-versicolor*
53	6.9	3.1	4.9	1.5	*Iris-versicolor*
54	5.5	2.3	4	1.3	*Iris-versicolor*
55	6.5	2.8	4.6	1.5	*Iris-versicolor*
56	5.7	2.8	4.5	1.3	*Iris-versicolor*
57	6.3	3.3	4.7	1.6	*Iris-versicolor*
58	4.9	2.4	3.3	1	*Iris-versicolor*
59	6.6	2.9	4.6	1.3	*Iris-versicolor*
60	5.2	2.7	3.9	1.4	*Iris-versicolor*
⋮					
101	6.3	3.3	6	2.5	*Iris-virginica*
102	5.8	2.7	5.1	1.9	*Iris-virginica*
103	7.1	3	5.9	2.1	*Iris-virginica*
104	6.3	2.9	5.6	1.8	*Iris-virginica*
105	6.5	3	5.8	2.2	*Iris-virginica*
106	7.6	3	6.6	2.1	*Iris-virginica*
107	4.9	2.5	4.5	1.7	*Iris-virginica*
108	7.3	2.9	6.3	1.8	*Iris-virginica*
109	6.7	2.5	5.8	1.8	*Iris-virginica*
110	7.2	3.6	6.1	2.5	*Iris-virginica*
⋮					

- *Knowing the physics of the problem helps understanding data.*
- *Iris is a genus of species of flowering plants with showy flowers. (In Arabic: Alsawsan).*
- *Iris is extensively grown as ornamental plant, medicine, drugs.*



Chapter 3

1-D charts

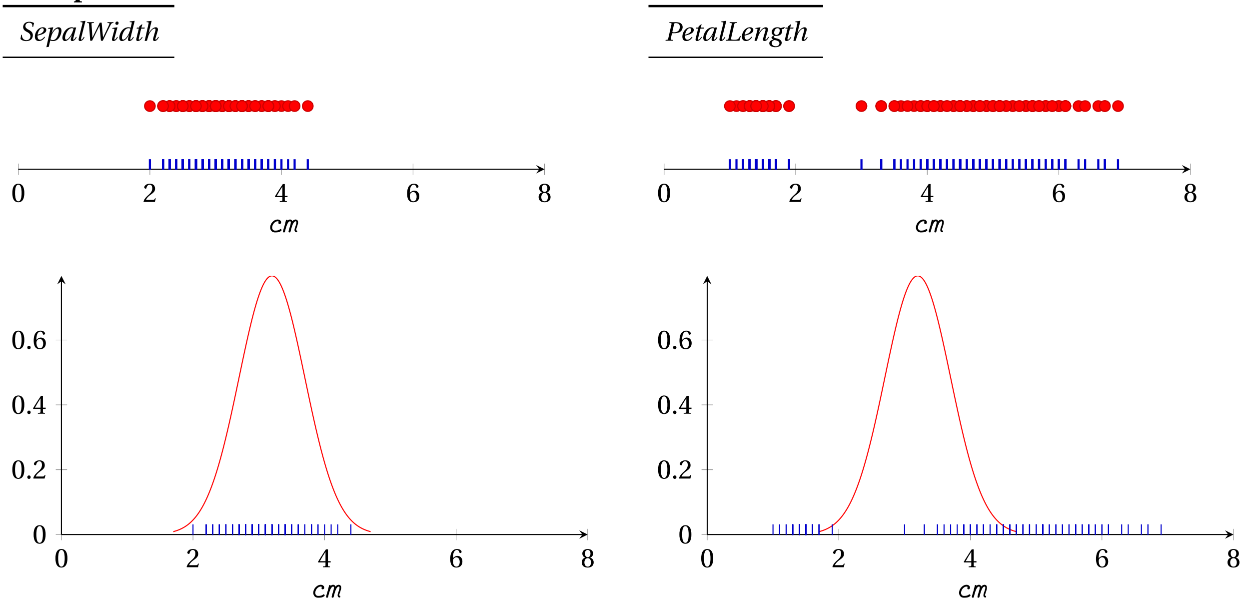
Per Hoaglin et al. (2000):

- How nearly symmetric the sample is?
- How spread out the numbers are?
- Whether a few values are far removed from the rest?
- Whether there are concentrations of data?
- Whether there are gaps in the data?

3.1 A Quantitative Variable

3.1.1 Rug Plot (the simplest ever)

Example 10 (`iris` **dataset)** .



Hints for sense: Some observations clutter each other; standardize scale,
@@@ use Gaussian fit (blindly before visualization) then Gaussian mixture (after visualization). Then calculate the probability of having 2<X<3 and compare to no.obs/150 .

@@@ for the red normal distribution, calculate mu and sigma from data.

3.1.2 Stem-and-Leaf

@@@ needs revisiting and drawing
Invented by by John W. Tukey, (who also coined the word **bi**nary digi**t**).
It is a multi-functioning of the "data measure", i.e., the displaying element (here the digit) has more than one function (position and value).
Variations: e.g., adding rank left to each stem.
$L = \lfloor 10 \times \log_{10} n \rfloor$ very good for $20 < n < 300$

3.1.3 Histograms: (for more details check St 121.)

$$I_{(c)} = \begin{cases} 1 & \text{if } c \text{ is } T \\ 0 & \text{if } c \text{ is } F \end{cases}, \qquad \text{(indicator function)}$$

$$I_{(c)} \sim Bernoulli\left(\Pr(c)\right).$$

For data $x_1, \ldots, x_n$ divide the data range T to K equal regions of equal width Δ (so that $K = T/\Delta$)

$$\begin{aligned} T_k &= [t_0 + \Delta k, t_0 + \Delta(k+1)[\\ &= [t_k, t_{k+1}[, \ k = 0, \ldots, K-1, \end{aligned}$$

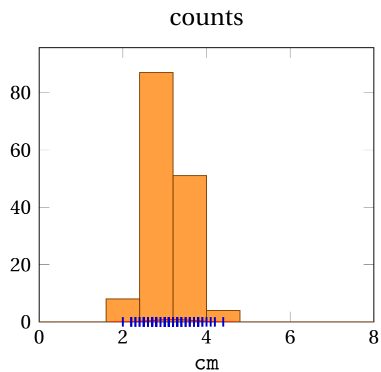
Notice: decreasing Δ increases K.
We have three versions of histogram:

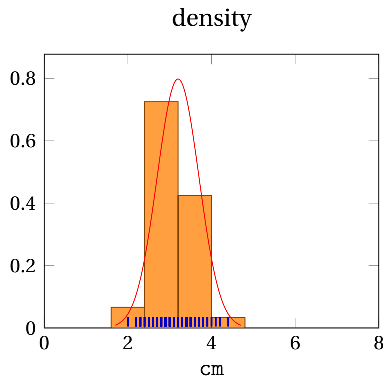
$$N_k = \sum_{i=1}^{n} I_{(X_i \in T_k)}, \qquad \text{(counts)}$$

$$R_k = \frac{N_k}{n} \xrightarrow{p} \Pr(X \in T_k) \qquad \text{(relative counts)}$$

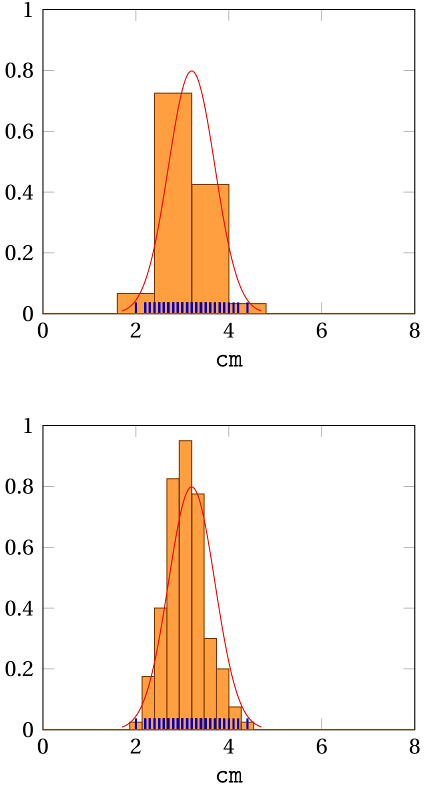
$$f_k = \frac{N_k}{\Delta n} \xrightarrow{p} \frac{\Pr(X \in T_k)}{\Delta} \approx \frac{f_X(t_k)\,\Delta}{\Delta} = f_X(t_k) \qquad \text{(density)}$$

SepalWidth

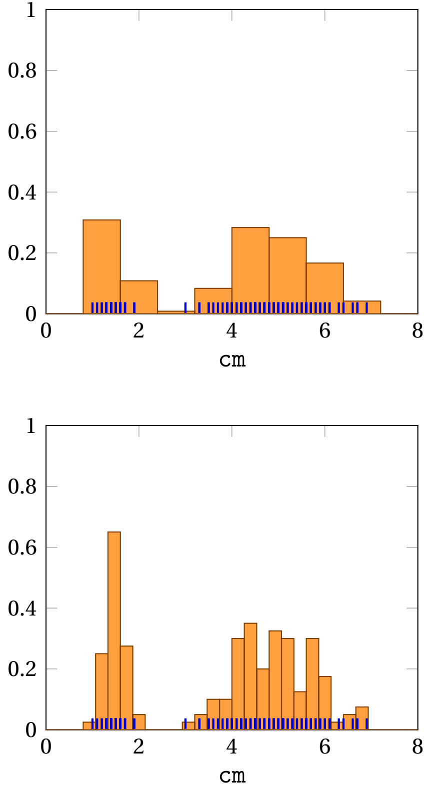




SepalWidth



PetalLength



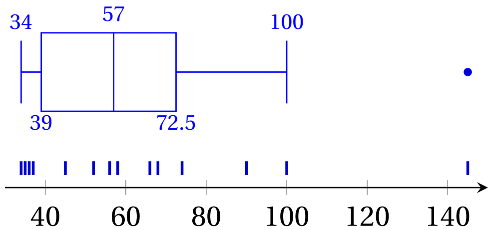
Hints for sense: bins = 10 vs. 30; unify X and Y scale for comparison;

3.1.4 Box Plot

back to example 11
To observe a glance: location, spread, skewness, tail length, and outlying data points.

lower whisker	lower quartile	median	upper quartile	upper whisker
$Q_L - 1.5d_Q \leq \min x_i = W_L$	Q_L	M	Q_U	$W_U = \max x_i \leq Q_U + 1.5d_Q$

Example 11 (Letter Values) .



Rank of M, QL, QU is 7.5, 4.25, 10.75

$M = 56 + 0.5(58 - 56) = 57$

$Q_L = 37 + 0.25(45 - 37) = 39$

$Q_U = 68 + 0.75(74 - 68) = 72.5$

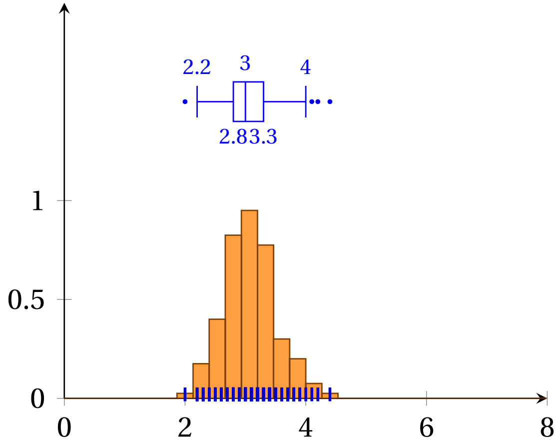
$d_Q = (72.5 - 39) = 33.5$

$39 - 1.5 \times 33.5 = -11.25$

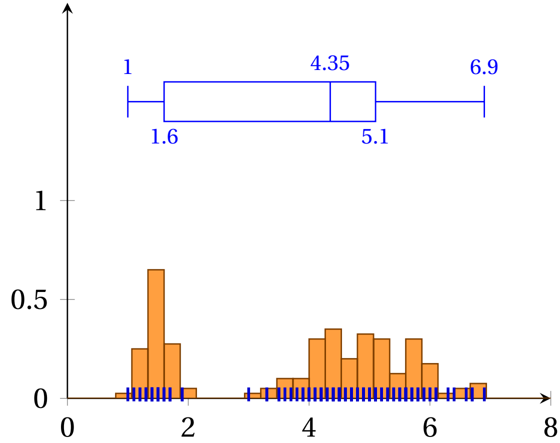
$72.5 + 1.5 \times 33.5 = 122.75$

we could have defined a boxplot based on mean and variance => less resistant.
Why boxplot is not defined in terms of $W_L = \widehat{F}^{-1}(0.05)$
why boxplot is not defined in terms of mean and variance
for small patches $\frac{\#\text{ of obs.} n}{>} \Pr(X < W_L)$

SepalWidth



PetalLength



Comparison

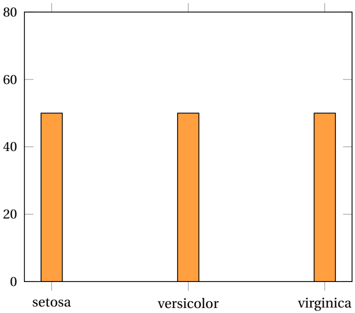
	Rug plot	**Histogram**	**Boxplot**	**Stem-and-leaf**
density	0 (clutter)	1	0 (region)	1
values	1	1	0 (region)	1
large N	0 (clutter)	1	1	0
resistance	0 (outliers)	0 (outliers)	1	0 (outliers)
discrete	0 (clutter)	1	1	1

Clutter could be alleviated by α-channel.

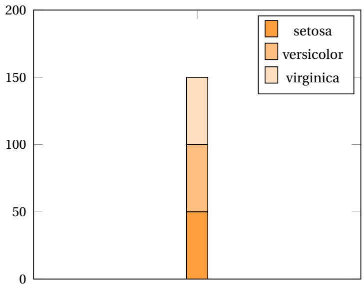
3.2 A Categorical Variable

Suppose we have only last column of table in Sec. 9; no numerical values. Only histogram-like charts: bar chart, stacked bar, pie chart, or any equivalent.

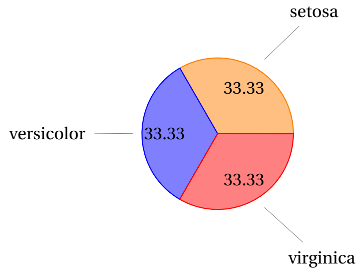
3.2.1 Bar chart



3.2.2 Stacked plot



3.2.3 Pie chart



- Bar chart is more professional and scientific; pie chart is more for illustration.
- More details can be put on the bar chart (including boxplot for each class, etc.)
- Bar chart and Stacked plot are utilized more for several patches.

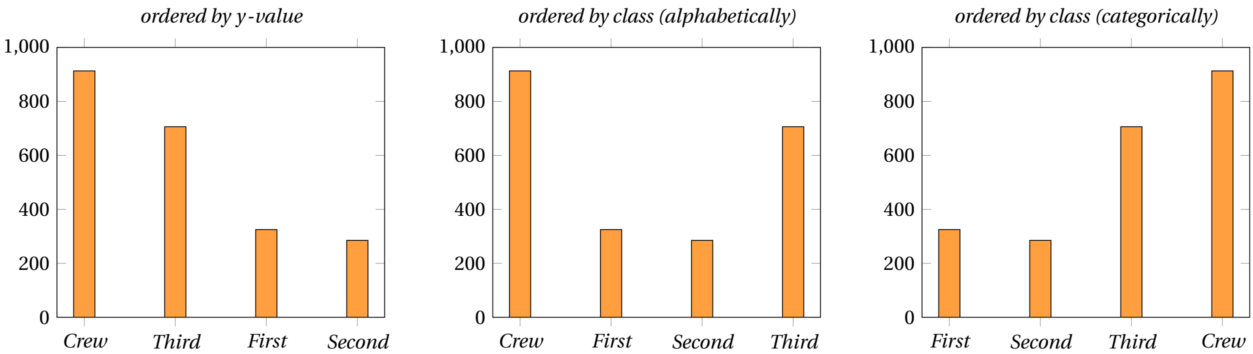
3.3 An Ordered Categorical Variable

Exactly as "bar chart" with ordered x-axis.

Example 12 (No. of Titanic passengers and crew) *: We can consider the variable (*`passenger class`*) as:*

- *categorical (as previous example) and order by y-value. (sorting will provide more information for the same ink).*
- *ordered categorical and order it alphabetically (nonsense in this example).*
- *ordered categorical and order it by class rank (makes sense here).*

Reproduced



Chapter 4

2-D Charts

This is the chapter

Part II

The Art of Visual Display, Presentation, and Illustration

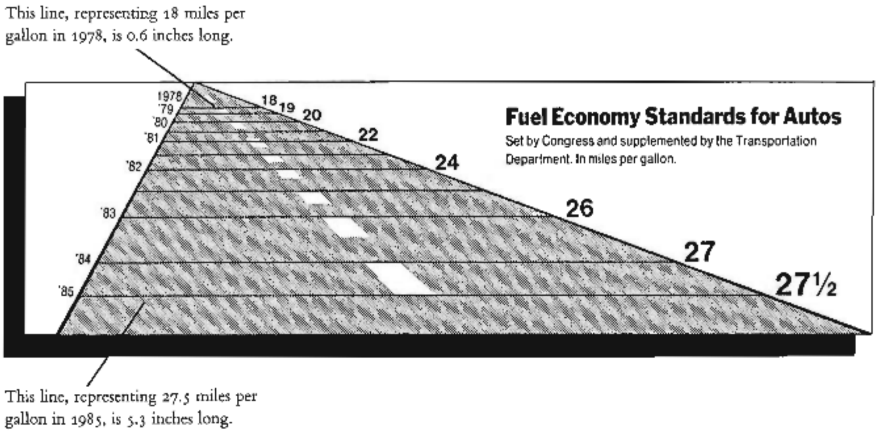
Chapter 5

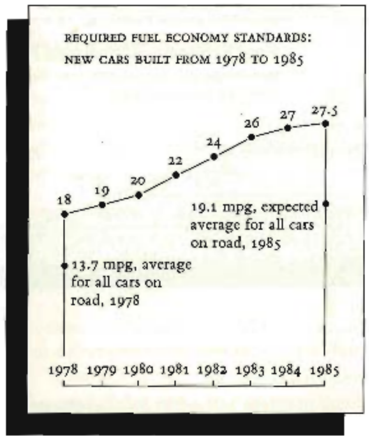
The Visual Display of Quantitative Information (Tufte, 2001)

@@@ data ink ratio before this example, because it has a problem in this ratio as well.

Example 13 (Lie Factor) *: (Tufte, 2001, P. 53) (The figure was published in New York Times, August 9, 1978)*

- *Three kinds of lies: lie, damn lie, and Statistics; also charts as well.*
- *Example of Statistics: Stock letters.*
- *Example of chart this one.*
- *Could have been decorated honestly like this one.*





$$\textbf{\textit{Lie Factor}} = \frac{\textit{size of effect shown in graphic}}{\textit{size of effect in data}} = \frac{(5.3-0.6)/0.6}{(27.5-18)/18} = \frac{7.83}{0.53} = 14.8$$

Part III

Applications

Bibliography

Chen, C.-h., Härdle, W., Unwin, A., 2008. Handbook of data visualization. Springer, Berlin.

Cook, D., Buja, A., Lang, D. T., Swayne, D. F., Hofmann, H., Wickham, H., Lawrence, M., 2007. Interactive and Dynamic Graphics for Data Analysis: With R and GGobi. Springer Science & Business Media.

Hoaglin, D. C., Mosteller, F., Tukey, J. W., 1985. Exploring data tables, trends, and shapes. Wiley, New York.

Hoaglin, D. C., Mosteller, F., Tukey, J. W., 2000. Understanding robust and exploratory data analysis, wiley clas Edition. Wiley, New York.

Tufte, E. R., 1990. Envisioning Information. Graphics Press, Cheshire, Conn.

Tufte, E. R., 1997. Visual explanations : images and quantities, evidence and narrative. Graphics Press, Cheshire, Conn.

Tufte, E. R., 2001. The visual display of quantitative information, 2nd Edition. Graphics Press, Cheshire, Conn.

Tufte, E. R., 2006. Beautiful evidence. Graphics Press, Cheshire, Conn.