Data Analysis
A single variable: shape and distribution

Prof. Dr. Jose Fernando Rodrigues Junior (v. original)
Profa. Maria Cristina

ICMC-USP

What is it about?

- Data analysis concerns **a better comprehension not of the data, but of the domain** that generated the data, so:

- **Start up with the problem domain**
  - For example: if it is a company, know its problems, its clients and products, its weaknesses and strong points, and **have a few analytical goals beforehand** – why did you start analyzing in the first place?

- Only after then, data analysis will lead you to **patterns and insights** that will foster: **prediction, evaluation, and determination of alternatives** → decision support

First hint – keep it simple

- Data analysis **sometimes gets more complicated than it should**, when one uses solutions way more complicated than what is necessary

- Many times**, the simplest techniques can do the job**
- Quoting:
  - **Simple** instead of complex
  - **Cheap** instead of expensive
  - **Explicit** instead of opaque
  - **Purpose** is more important than process
  - **Insight** is more important than precision
  - **Understanding** is more important than technique
  - **Think more**, work less

First hint – keep it simple

➔ Data analysis **sometimes gets more complicated that it should**, when one uses solutions way more complicated than

➔Common mistakes:

• Use of **statistical methods that you do not fully understand**

• Use of **complex black-box solutions**, when simple and transparent ones would work better

➔Risks

• **Simplicity is less prone to go wrong** without you noticing it

• **Complexity may obscure the obvious**

First: look into the data

- Part I of the book
  - Graphical analysis

  - Single variable – univariate
  - Two variables – bivariate
  - Time as variable – time series
  - More than two variables – multivariate data

Single variable: shape and distribution

- Questions on univariate data
  - **Size** of the data set
  - **Typical** and **extreme** (minimum and maximum) values
  - **Distribution** and its characteristics (symmetry, pareto-like,...)
  - **Clusters** and their characteristics (number, size, intersection...)
  - **Outliers**
  - Anything else that draws **attention**

Parentehesis: variable types

- A *categorical* variable, also called a *nominal* variable, is for mutual exclusive, but not ordered, categories.
  - For example, your study might compare five different genotypes. You can code the five genotypes with numbers if you want, but the order is arbitrary and any calculations (for example, computing an average) would be meaningless.

- A *ordinal* variable, is one where the order matters but not the difference between values
  - For example, you might ask patients to express the amount of pain they are feeling on a scale of 1 to 10. A score of 7 means more pain that a score of 5, and that is more than a score of 3. But the difference between the 7 and the 5 may not be the same as that between 5 and 3. The values simply express an order. Another example would be movie ratings, from * to *****.
- See http://www.graphpad.com/support/faqid/1089/

Parentehesis: variable types

- A *interval* variable is a measurement where the difference between two values is meaningful
  - The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees.
- A *ratio* variable, has all the properties of an interval variable, and also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable
  - Variables like height, weight, enzyme activity are ratio variables.
  - Temperature, expressed in F or C, is not a ratio variable. A temperature of 0.0 on either of those scales does not mean 'no heat'.
  - However, temperature in Kelvin is a ratio variable, as 0.0 Kelvin really does mean 'no heat'.
  - When working with ratio variables, but not interval variables, you can look at the ratio of two measurements. A weight of 4 grams is twice a weight of 2 grams, because weight is a ratio variable. A temperature of 100 degrees C is not twice as hot as 50 degrees C, because temperature C is not a ratio variable.
- See http://www.graphpad.com/support/faqid/1089/

Parentehesis: variable types

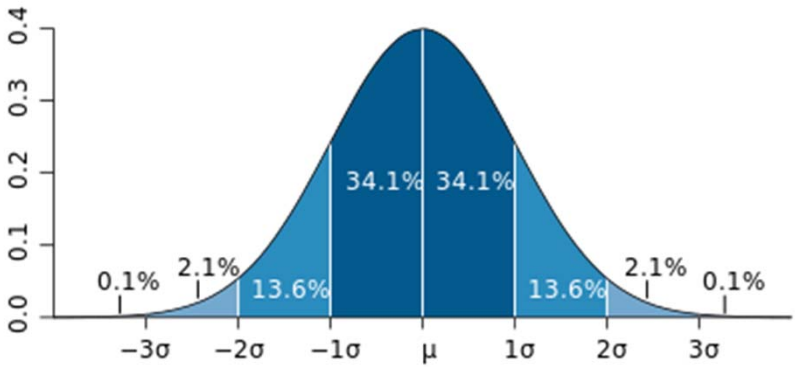
OK to compute...	**Nominal**	**Ordinal**	**Interval**	**Ratio**
frequency distribution	Yes	Yes	Yes	Yes
median and percentiles	No	Yes	Yes	Yes
add or subtract	No	No	Yes	Yes
mean, standard deviation, standard error of the mean	No	No	Yes	Yes
ratio, or coefficient of variation	No	No	No	Yes

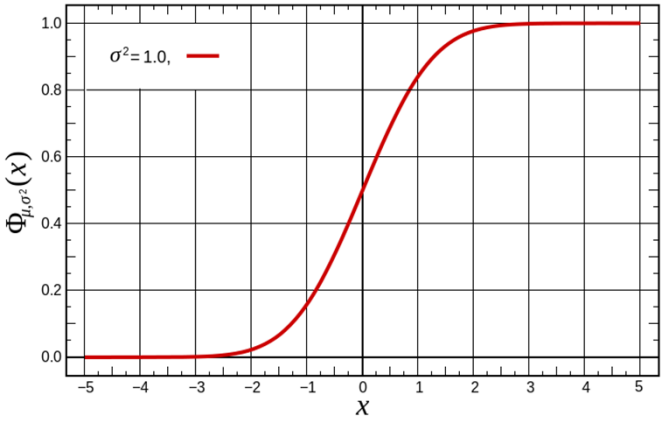
Common approach

- Summarize the data
  - Summary statistics: mean, variance, standard deviation...
  - Not as a starting point... even if the computations can be done

Pitfalls on basic summary statistics

- The most used summary statistics, **mean (μ), variance (σ^2 or s^2), and standard deviation (σ or s)**, although easy to use, only make sense for **distributions with a single peak (unimodal),** that is, **nearly symmetric without outstanding outliers**





Pitfalls on basic summary statistics

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2}$$

Pitfalls on basic summary statistics

- The most used summary statistics, **mean (μ), variance (σ^2 or s^2), and standard deviation (σ or s)**, although easy to use, only make s… is, r…

  ➔Mean is also named "**expected value**" in the sense that it is **the most common value to occur**

  ➔It is not hard to see **that multi-peaked distributions do not have the mean as the most common value**

  ➔**It these cases, basic summary statistics do no apply**

Pitfalls on basic summary statistics

- For example: suppose the prices of 11 items are **{1,1,1.5,0.5,1,1,1,1,1,1,20}**, the **mean price is $2.73** and the **standard deviation is $5.46**

- However, none of the items costs near that value;
- The implication that most of the items should cost between **$2.73-$5.46 and $2.73+$5.46** suggests items with absurd negative values

Median and the inter-quartile range (IQR)

- For situations like this, it is better to use the **median and the inter-quartile range (IQR)** in order to summarize the distribution of the data

- Quartiles: values that divide the data values (events) into 4 intervals of equal size
- Lower quartile: 25% of the events are lower than this value
- Median: 50% of the values are lower than this value (i.e., splits the data values into two intervals of equal size)
- Upper quartile: 75% of the events are lower than this value
- And the **inter-quartile range** (IQR) corresponds to the distance between the **Upper** and **Lower quartiles**, or the region where 50% of the events occur

Pitfalls on basic summary statistics

- **Example – Upper and lower quartiles**

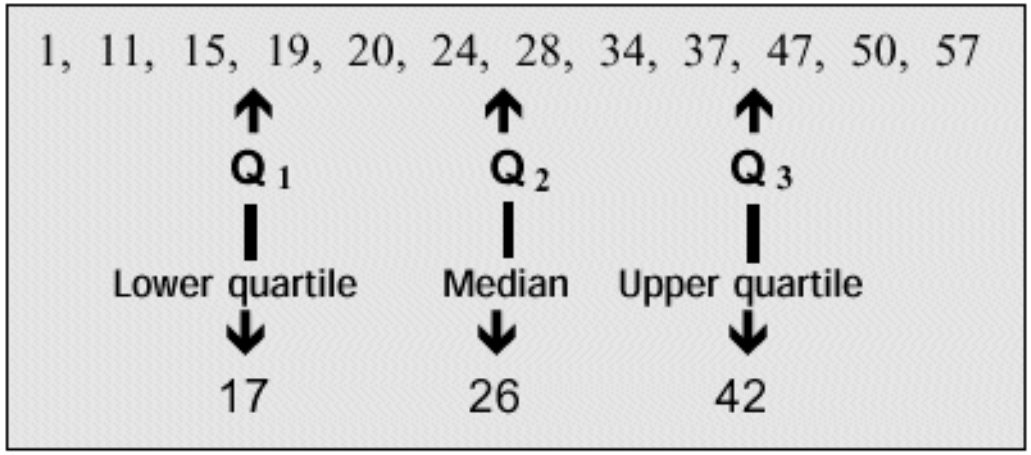
  - Data: 6, 47, 49, 15, 43, 41, 7, 39, 43, 41, 36
  - Ordered data: 6, 7, 15, 36, 39, 41, 41, 43, 43, 47, 49

  - Median (Q2): 41
  - Upper quartile (Q3): 43
  - Lower quartile (Q1): 15

See: http://www.statcan.gc.ca/edu/power-pouvoir/ch12/5214890-eng.htm

Pitfalls on basic summary statistics

- **Example 2 – Upper and lower quartiles**



See: http://www.statcan.gc.ca/edu/power-pouvoir/ch12/5214890-eng.htm

Median and the inter-quartile range (IQR)

- In general, the **p^{th} percentile** corresponds to **the smallest value x for which the cumulative distribution function, cdf(x) ≥ p**
- So, the **median** corresponds to the **50^{th} percentile**
- And the **inter-quartile** corresponds to the distance between the **75^{th} and the 25^{th} percentiles**, where 50% of the events shall occur
- This kind of summarization is commonly used for expressing the average salary in official publications; the median is used, not the mean; **the mean would be significantly distorted by the few households with extremely high incomes**

Median and the inter-quartile range (IQR)

- Median and IQR can always substitute mean and standard deviation; historically, they have not been preferred because **they are more computationally expensive** $O(n^2)$ against $O(n)$, due to the need of sorting

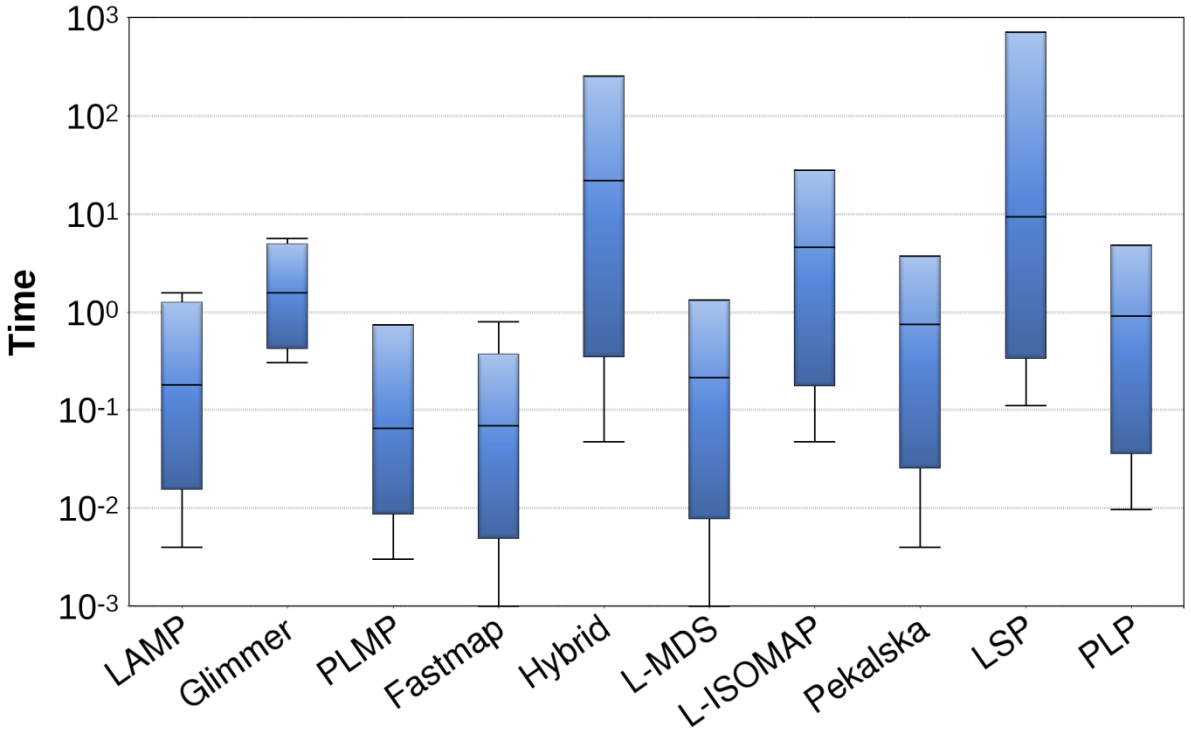
Pitfalls on basic summary statistics

- The median is conventionally defined as the value of a data set such that half of all points in the data set are smaller and the other half greater that that value. Percentiles are the generalization of this concept to other fractions (the 10th percentile is the value such that 10 percent of all points in the data set are smaller than it, and so on). Quantiles are similar to percentiles, only that they are taken with respect to the fraction of points, not the percentage of points (in other words, the 10th percentile equals the 0.1 quantile).

Conveying data distributions visually

- Box-plot charts
- Histograms
- Kernel density estimates
- Pareto charts

Box-Plots (Box-and-Whisker Plots)



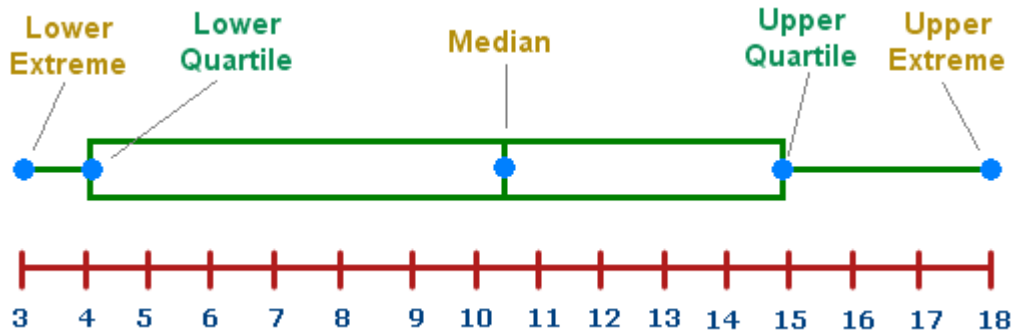
Fonte: Técnicas de projeção para identificação de grupos e busca por similaridade em dados multidimensionais. Tese de doutorado, P. Jóia Filho, 2015.

Box-Plots (Box-and-Whisker Plots)

- Box and whisker plots are ideal for comparing distributions because the centre, spread and overall range are immediately apparent
- A box and whisker plot is a way of summarizing a set of data measured on an *interval scale*. It is often used in explanatory data analysis. This type of graph is used to show the shape of the distribution, its central value, and its variability

Box plots (Box-and-Whisker Plots)

- A box plot can adequately express median and IQR
- It consists of:
  - A marker or symbol for the median – the location of the distribution
  - A box spanning the inter-quartile range – the width of the distribution
  - A set of whiskers extending from the center to the upper and lower extremes – the tails of the distribution



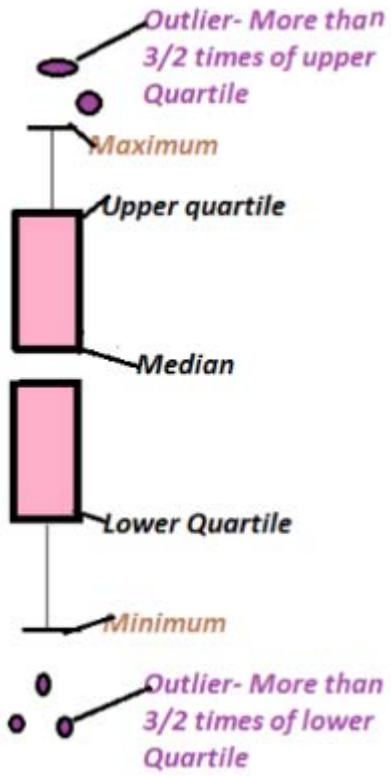
Source figure: http://www.mathcaptain.com/statistics/box-and-whisker-plot.html

Box plots (Box-and-Whisker Plots)

- A box plot can adequately express median and IQR
- It consists of:
  - A marker or symbol for the median – the location of the distribution
  - A box spanning the inter-quartile range – the width of the distribution
  - A set of whiskers extending from the center to the upper and lower *adjacent values* – the tails of the distribution
  - Individual symbols for outliers outside the adjacent ranges

- Lower adjacent value: values < (1,5 x IQR) + Q1
- Upper adjacent value: values > (1,5 x IQR) + Q3

Box plots (Box-and-Whisker Plots)



Source figure: http://www.mathcaptain.com/statistics/box-and-whisker-plot.html

Box plots (Box-and-Whisker Plots)

- A box plot can adequately express median and IQR
- It consists of:
  - A marker or symbol for the median – the location of the distribution
  - A box spanning the inter-quartile range – the width of the distribution
  - A set of whiskers extending from the center to the upper and lower *adjacent values* – the tails of the distribution
  - Individual symbols for outliers outside the adjacent ranges

- Lower adjacent value: values < (2 x IQR) + Q2
- Upper adjacent value: values > (2 x IQR) + Q2
- There is not a universal convention for box plots, this is only an usual one (book Janert)

Box plots (Box-and-Whisker Plots)

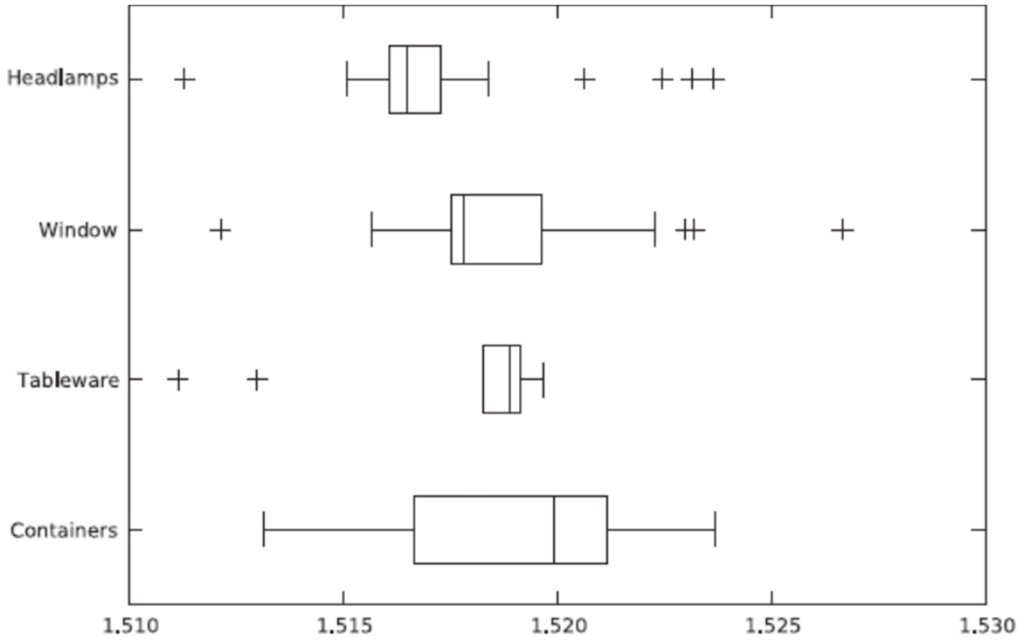
- A box plot can adequately express median and IQR
- It consists of:
  - A marker or symbol for the median – the location of the distribution
  - A box spanning the inter-quartile range – the width of the distribution
  - A set of whiskers extending from the center to the upper and lower adjacent values – the tails of the distribution
  - Individual symbols for outliers outside the adjacent ranges

where the upper adjacent value is the largest value in the data set that is less than twice the inter-quartile range greater than the median; the same for the lower

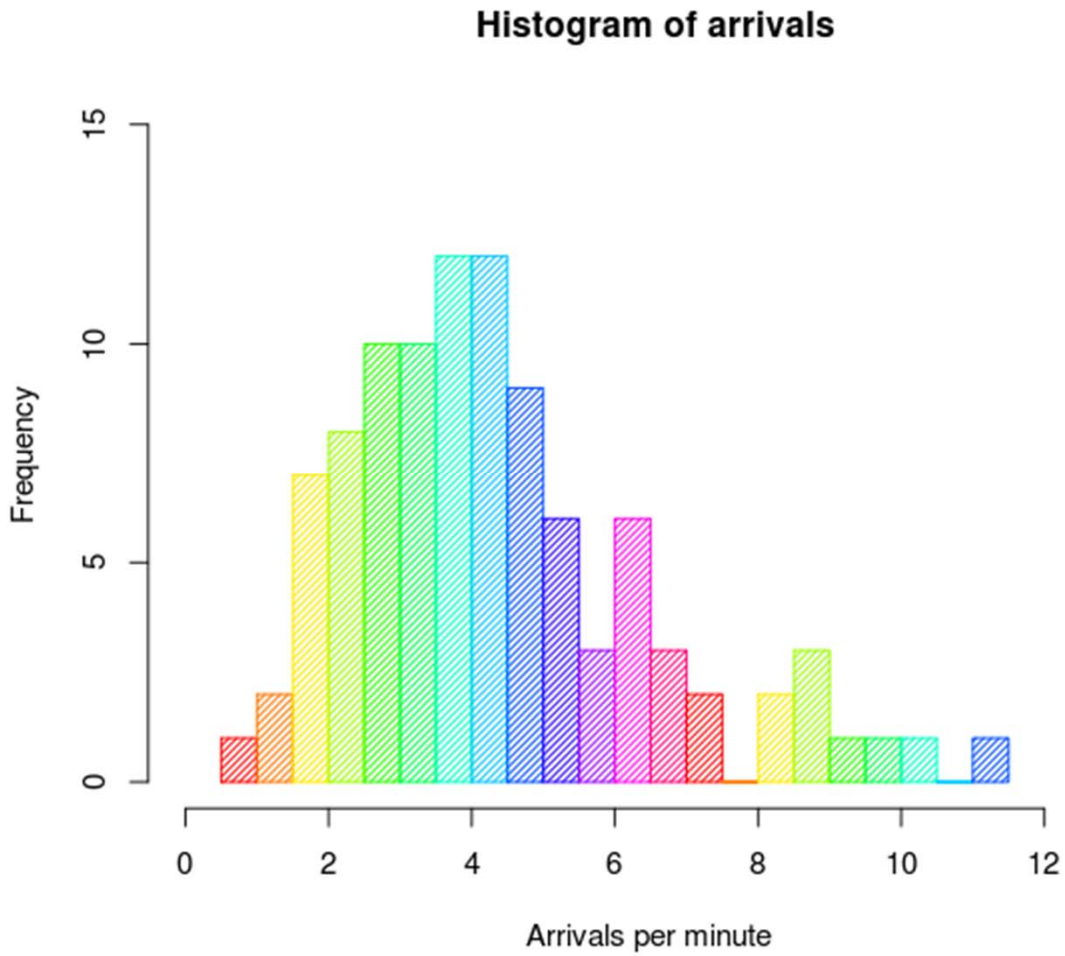
- There is not a universal convention for box plots, this is only an usual one

Box plots (Box-and-Whisker Plots)

- For example, consider a data set containing the index of refraction of 121 samples of glass; the data set is broken down by the type of glass: 70 samples of window glass, 29 from headlamps, 13 from containers of various kinds, and 9 from tableware



Histograms



https://en.wikipedia.org/wiki/Histogram

Histograms

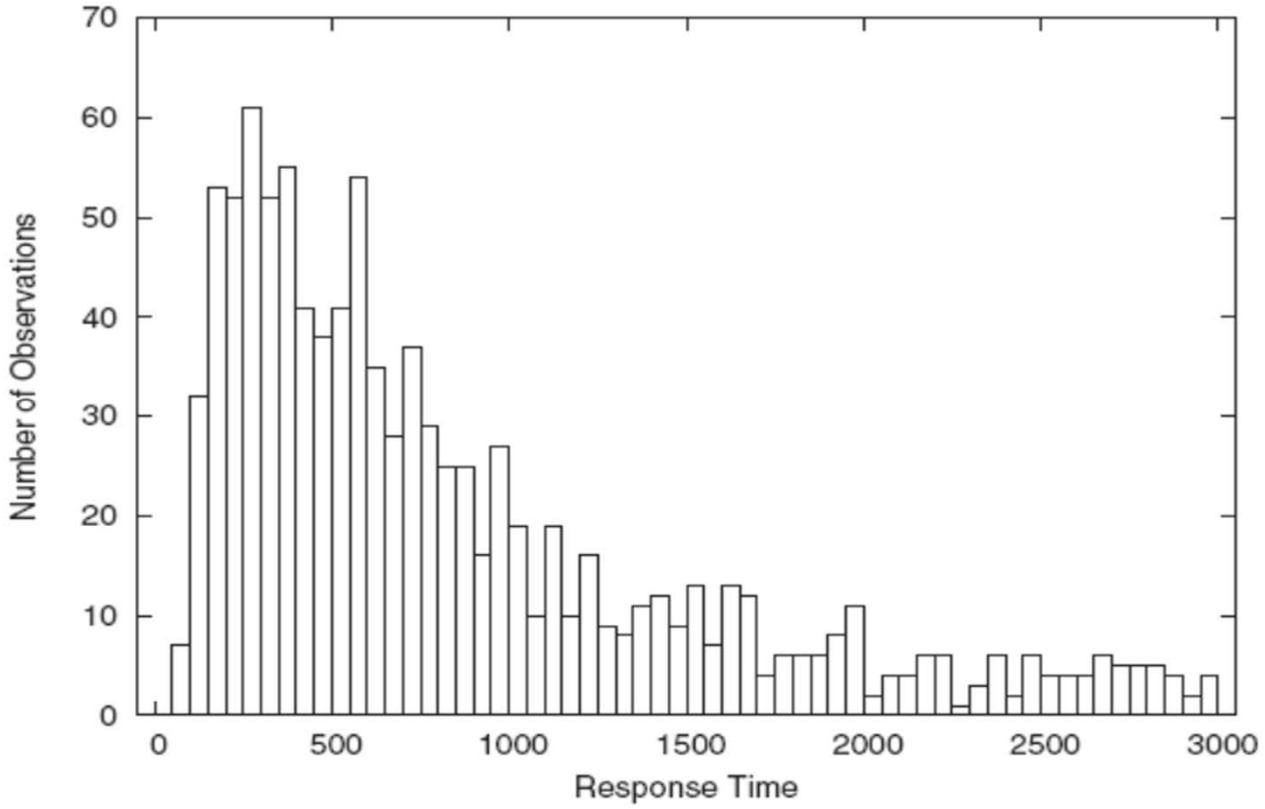
- **How-to:**
  - **Divide the range of values into a set of bins**
  - **Count the number of points that fall into each bin**
  - **Plot the count as a function of the position of each bin**
- Example:
  - Suppose a dataset with 1,000 data points, each one corresponding to the time a web server takes to answer to a given service request:

- 452.42
- 318.58
- 144.82
- 129.13
- 1216.45
- 991.56
- 1476.69
- 662.73
- 1302.85
- 1278.55
- 627.65
- 1030.78
- 215.23
- 44.50
- ...

Histograms

- **How-to:**
  - **Divide the range of values into a set of bins**
  - **Count the number of points that fall into each bin**
  - **Plot the count as a function of the position of each bin**
- Example:
  - Suppose a dataset with 1,000 data points, each one corresponding to the time a web server takes to answer to a given service request:

- 452.42
- 318.58
- 144.82
- 129.13
- 1216.45
- 991.56
- 1476.69
- 662.73
- 1302.85
- 1278.55
- 627.65
- 1030.78
- 215.23
- 44.50
- ...



Histograms

- What does the histogram tell us?
  - A sharp cutoff at the left, pointing out the minimum time to complete a request
  - A sharp rise to a maximum – the typical response time
  - A reasonable large tail to the right – the requests that take longer to complete
- This is a **typical histogram** for task-completion times, as for students in homework, or manufacturing workers: **a minimum time no one can beat, a small group of faster champions, a large majority, a tail of stragglers**

Histograms – problems

- Many histograms can be produced from a single data set
  - **bin width**
  - **bin alignment**

Histograms – problems

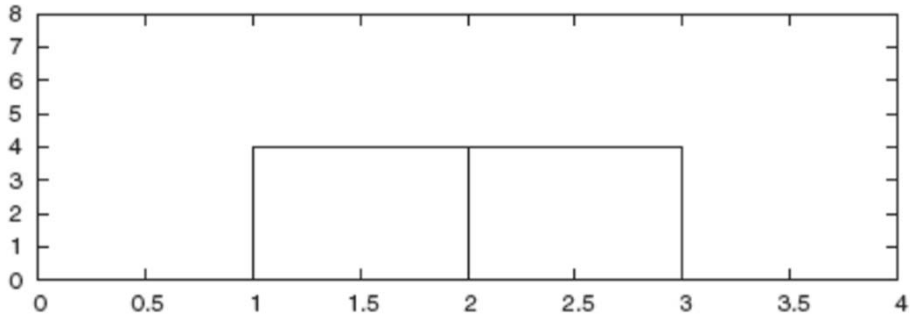
- How to determine the **bin width**?
  - In the previous example it was set to 50ms, but **it could be any other value**
  - If the bin is **too large, you may loose details**; if it is **too small, you loose representativeness**
  - Solution: **trial and error** or, for **Gaussian distributions**, use Scott's rule with the bin width $\mathbf{w = 3.5\sigma/\sqrt[3]{n}}$, for σ is the standard deviation and n is the number of points

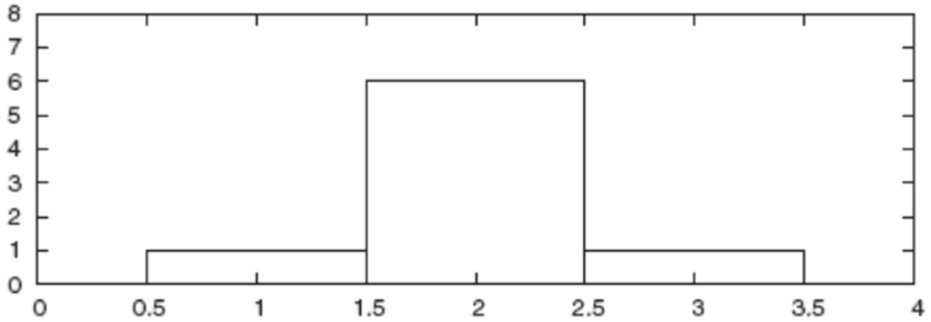
Histograms – problems

- How to **align the bins**?
  - In the previous example, the first bin was set to start at multiples of 50 (1-50,51-100,101,150, …) but it could be (25-75,76-125, …), for example
  - **Together with the bin width, the bin alignment defines which elements fall into each bin, totally redefining the histogram appearance**

Histograms – problems

- How to **align the bins**?
  - For example, consider the data set {1.4, 1.7, 1.8, 1.9, 2.1, 2.2, 2.3, 2.6} and histograms with width 1 and alignment set by multiples of 1 and 0.5, respectively:





Histograms – problems

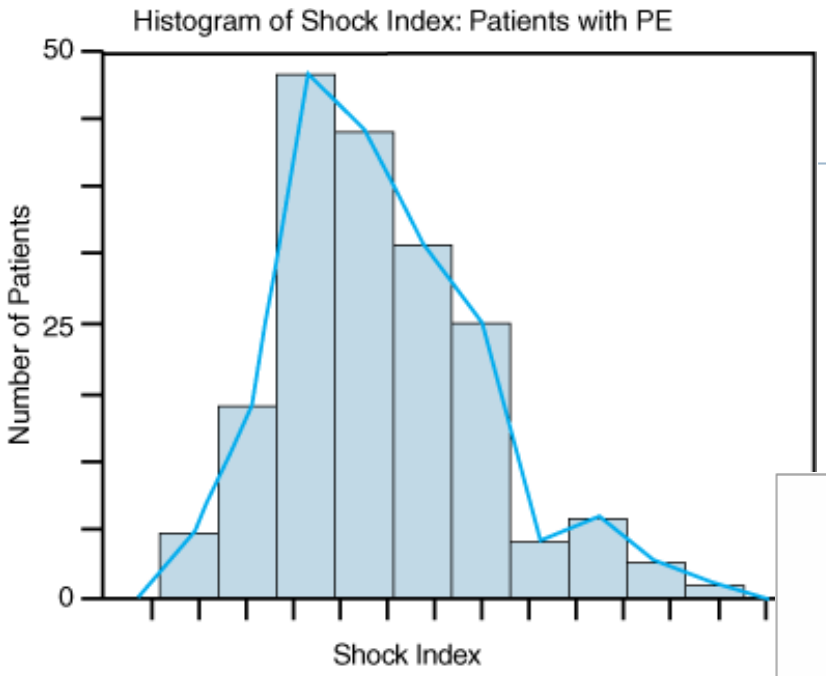
➔These two parameters can totally redefine the histogram, **for bad or for good**

➔As in the example, the histograms lend two interpretations of the data, **which one is correct depends on the analytical goals, on the data unit, on the familiarity of the analyst, on the usual practice, and on the very appearance of the histogram**

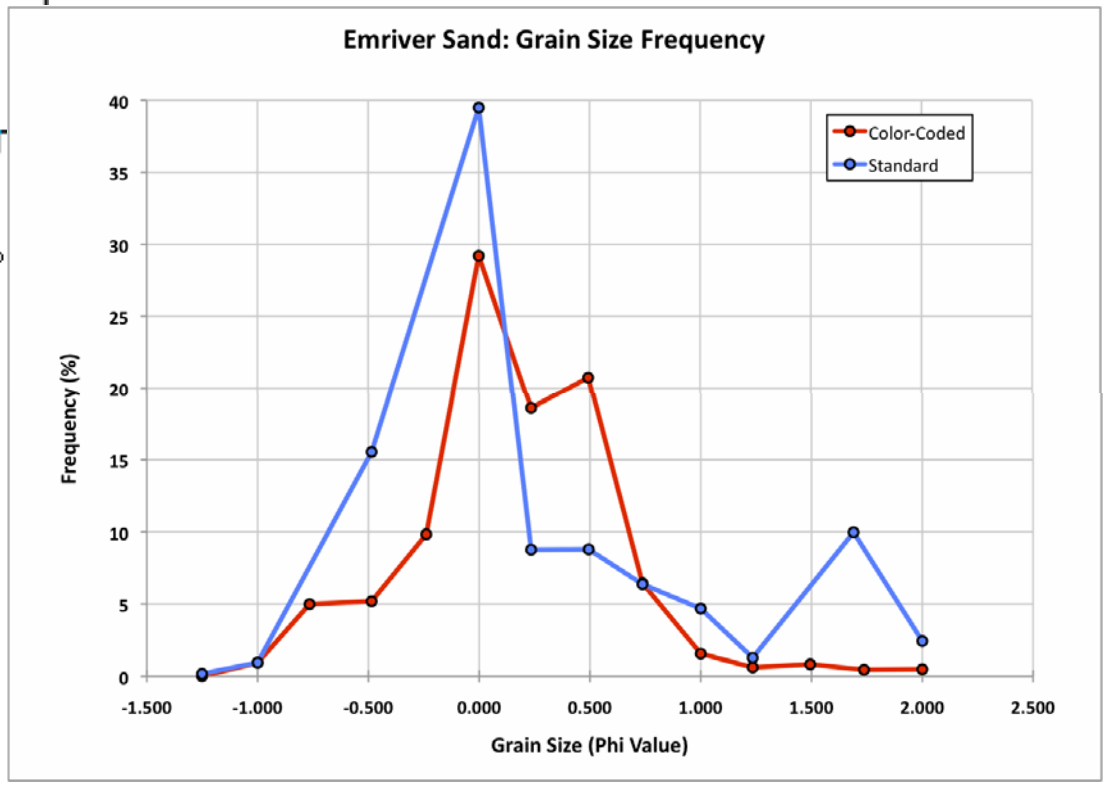
➔Unfortunately, **there is no easy way** to determine what the best layout is

Histograms – other issues

- Histograms can be **normalized (percentage)** or **not (absolute values)**
- Bins can have **variable width**, despite being potentially ambiguous, which is useful for unbalanced sets
- **Multiple data sets** can occupy the same histogram, however not using rectangles (use frequency polygons instead)



Source: Dawson B, Trapp RG: *Basic & Clinical Biostatistics*, 4th Editio
http://www.accessmedicine.com



Histograms – other issues

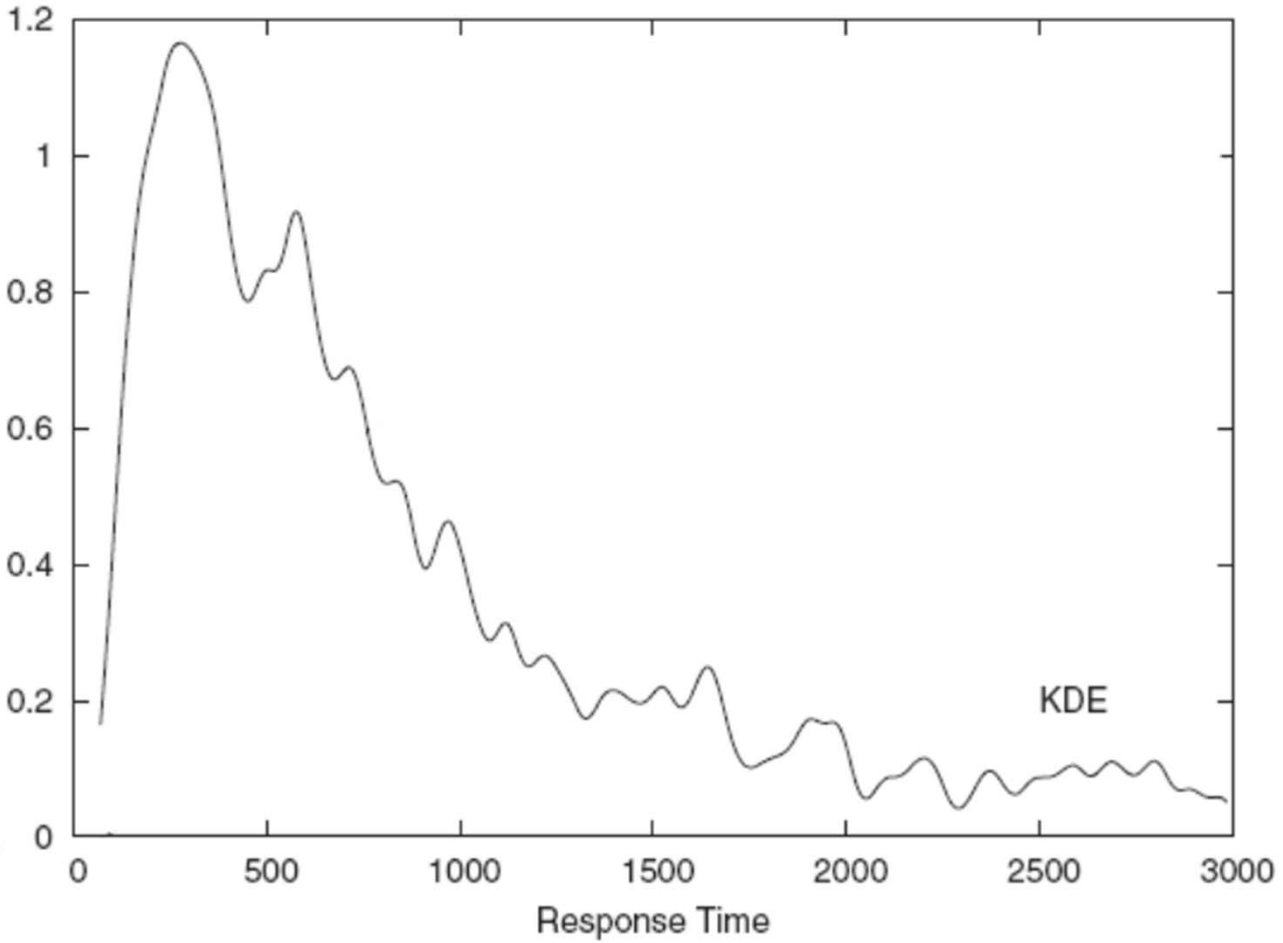
➔ A similar, though more robust approach, is to use Kernel Density Estimates

Kernel Density Estimates (KDE)

- **How-to:**
  - Place a kernel (smooth, strongly peaked function) at the position of each data point
  - Add up the contributions of each kernel to obtain a smooth curve
- **Advantages over histograms**
  - Continuous, smaller loss of information
  - Smooth rather than ragged plotting
- **Disadvantage**
  - More computationally costly than histograms

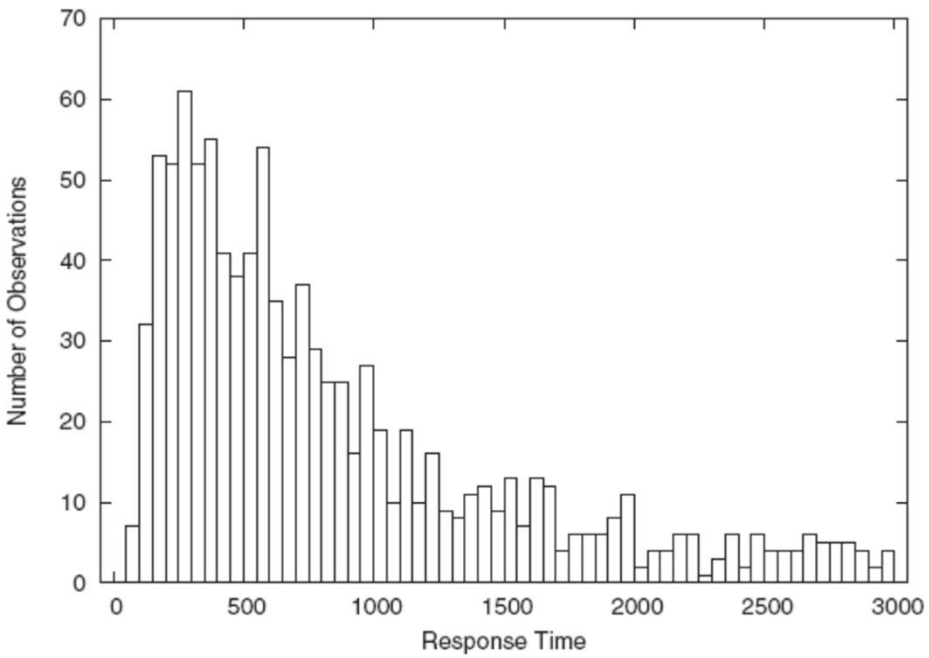
Kernel Density Estimates (KDE)

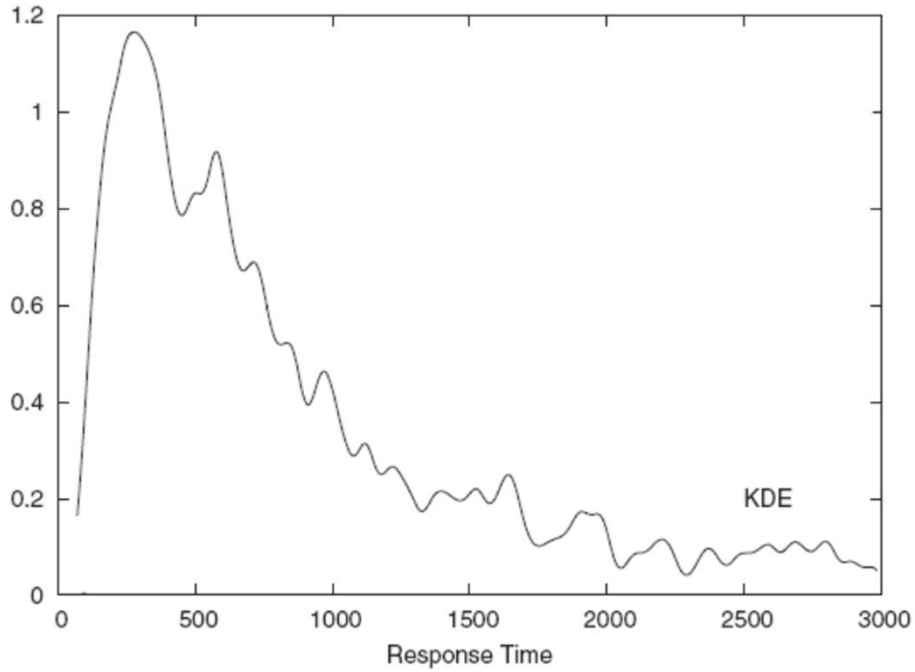
- **Example**: the web server time response data (slide 32) now presented with KDE



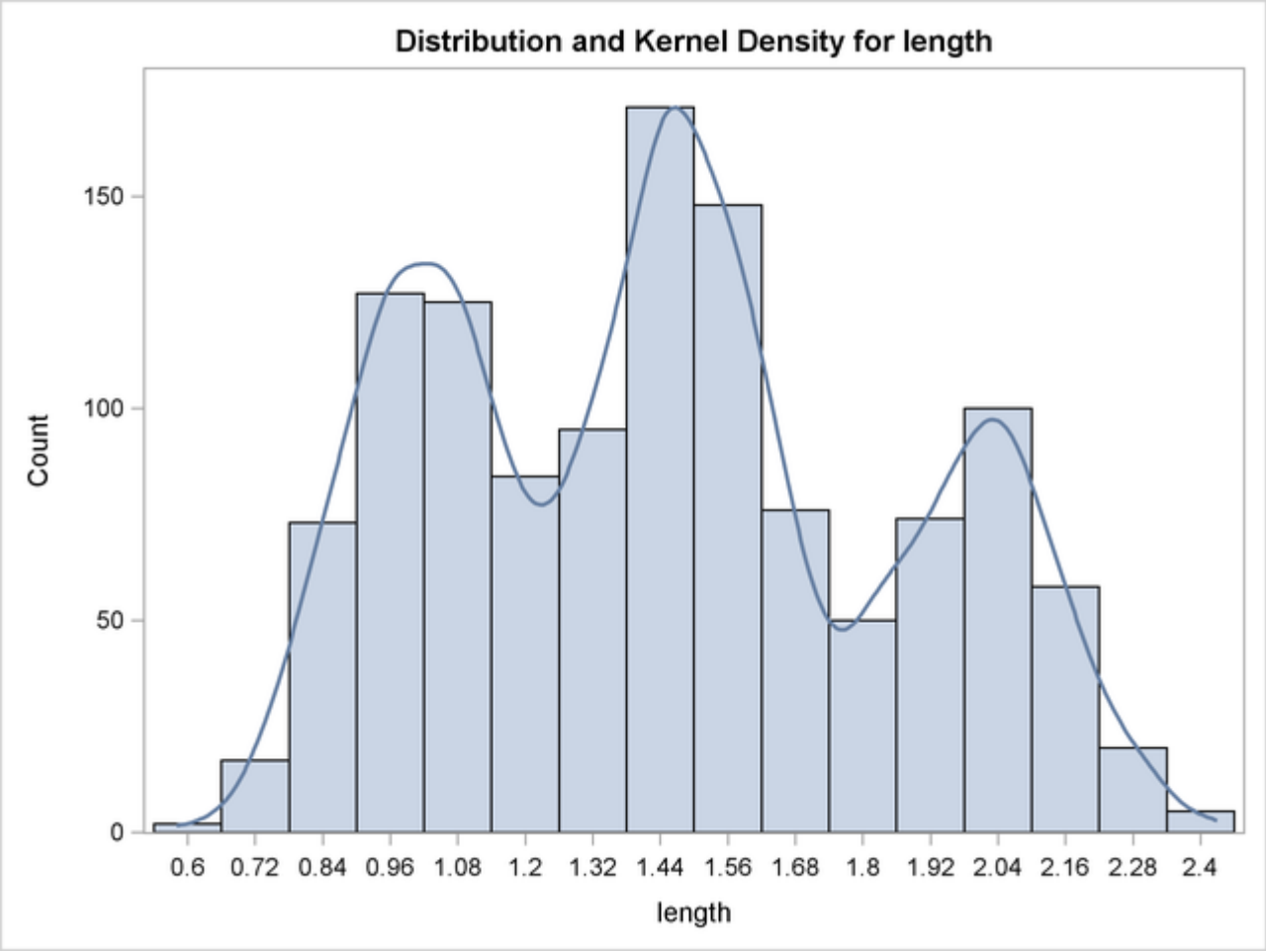
Kernel Density Estimates (KDE)

- **Example**: the web server time response data (slide 6) now presented with KDE





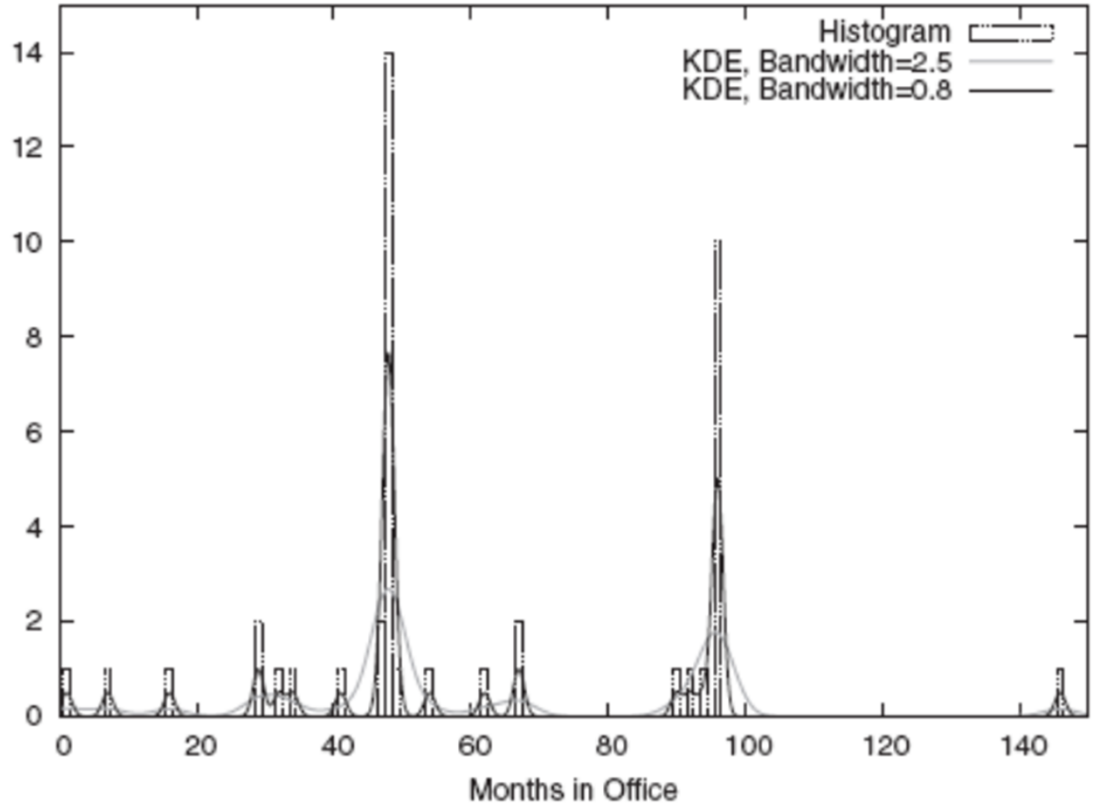
Histogram + KDE



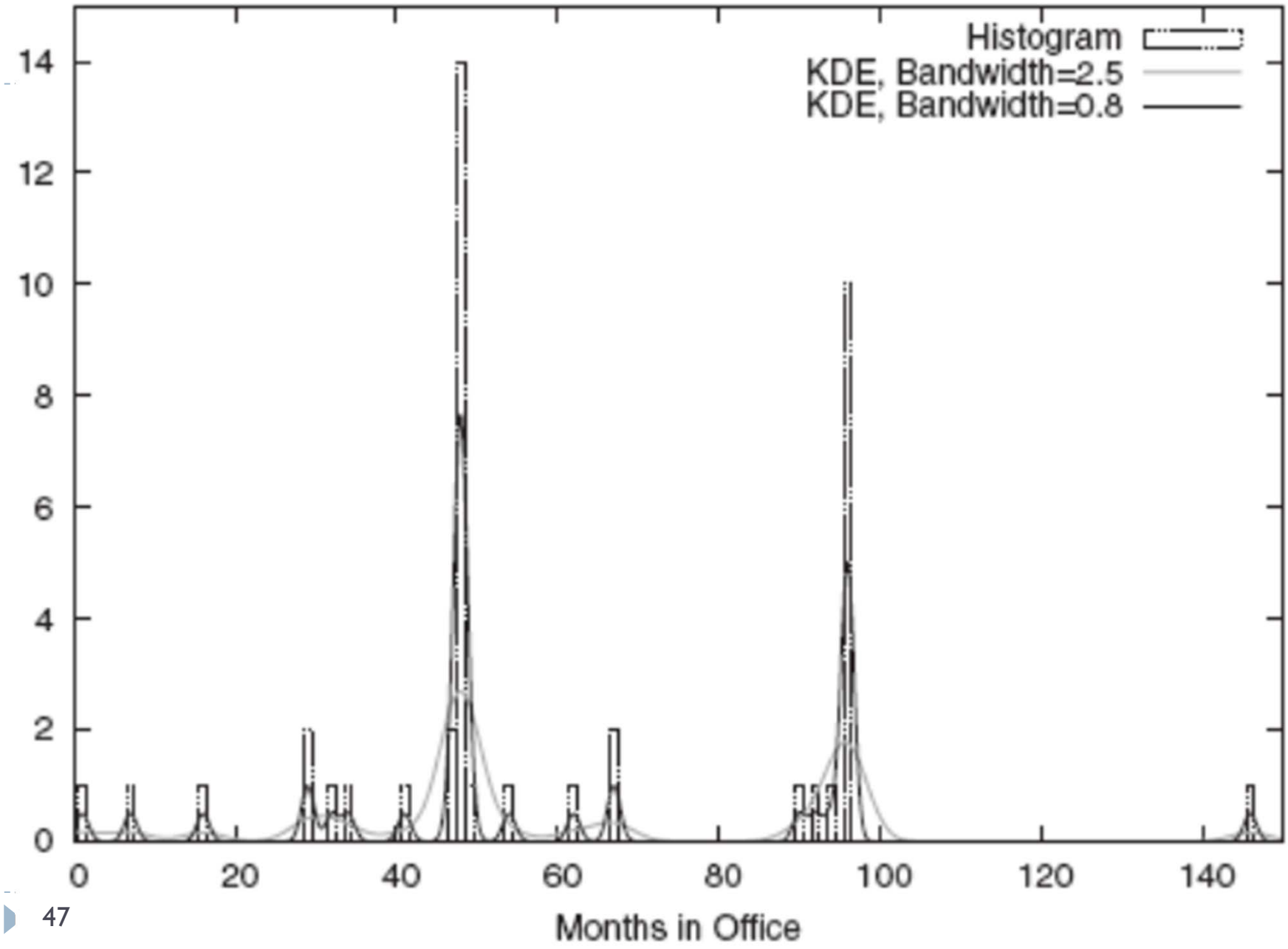
Kernel Density Estimates (KDE)

- **Another example**: consider the following data set

Id - USA President – Number of days in office
1 Washington 94
2 Adams 48
3 Jefferson 96
4 Madison 96
5 Monroe 96
6 Adams 48
…
39 Carter 48
40 Reagan 96
41 Bush 48
42 Clinton 96
43 Bush 96



Extracted from *http://www.amstat.org/publications/jse/v13n1/datasets.hayden.html*


Histogram
KDE, Bandwidth=2.5
KDE, Bandwidth=0.8
14
12
10
8
6
4
2
0
0
20
40
60
80
100
120
140
Months in Office

Kernel Density Estimates – a little math

- In KDE, kernels must be **functions that integrate to 1**, as for example

box or boxcar

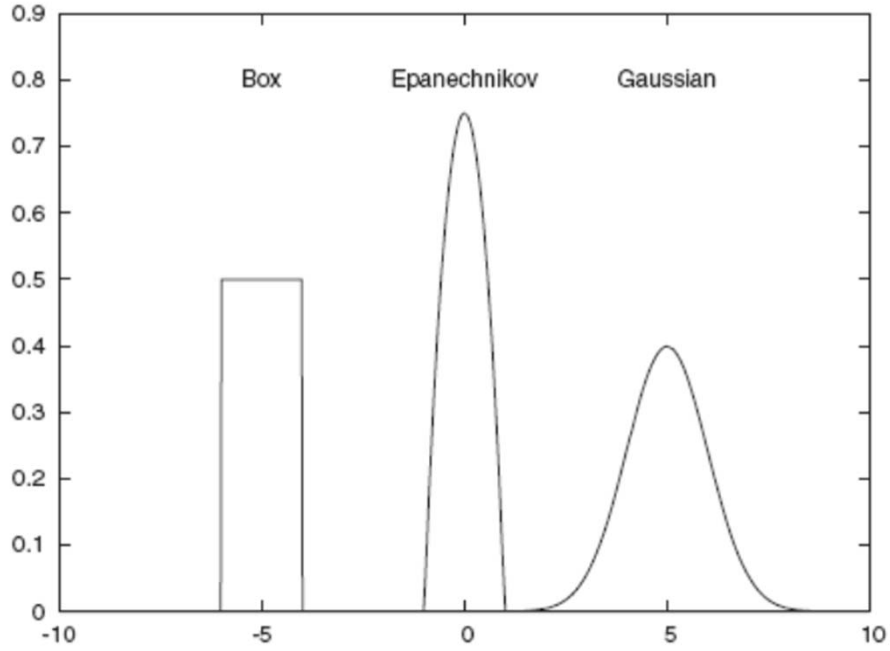
$$K(x) = \begin{cases} \frac{1}{2} & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Epanechnikov

$$K(x) = \begin{cases} \frac{3}{4}\left(1 - x^2\right) & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Gaussian

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$



Kernel Density Estimates – a little math

- The most commonly used function is the **Gaussian**
- When using it, the first thing is to **shift the kernel** to the position of each point, so instead of K(x), we **use K(x-x_i)**, so that its **peak will be at x_i, not at 0**
- It is also necessary to choose the kernel **bandwidth h** and set it so that the area under the curve remains 1, it goes **as follows**:

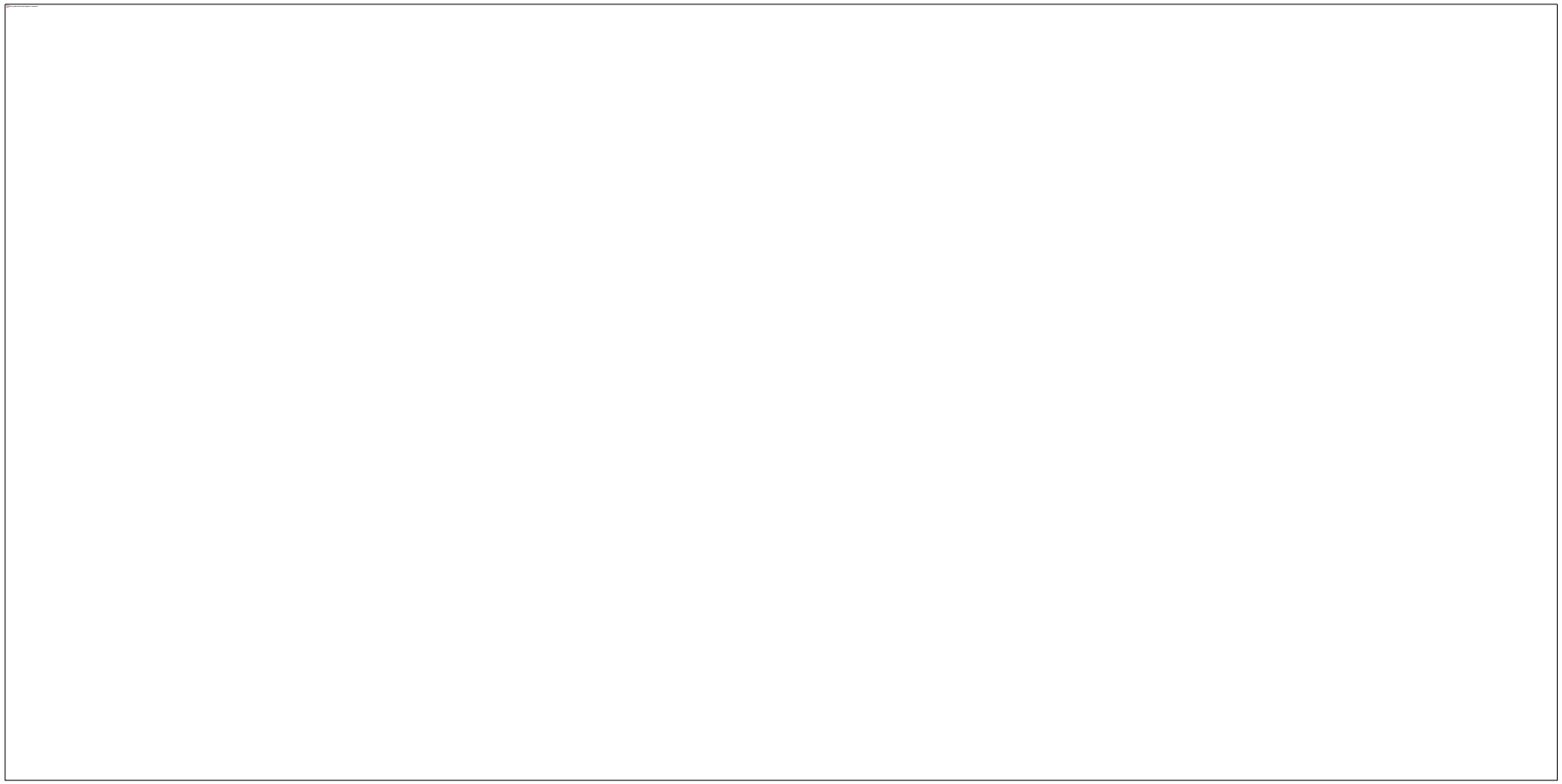
$$\frac{1}{h}K\left(\frac{x-xi}{h}\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$$

Kernel Density Estimates – a little math

- Finally, for a data set $\{x_1, x_2, x_3, \ldots, x_n\}$, the smooth KDE curve is given by:

$$D_h(x; \{x_i\}) = \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - xi}{h}\right)$$

Kernel Density Estimates (KDE)



Comparison of the histogram (left) and kernel density estimate (right) constructed using the same data. The data points are the rug plot on the horizontal axis.
https://en.wikipedia.org/wiki/Kernel_density_estimation

Kernel Density Estimates (KDE)

- **Another example**: consider the following data set

Id - USA President – Number of days in office
1 Washington 94
2 Adams 48
3 Jefferson 96
4 Madison 96
5 Monroe 96
6 Adams 48
…
39 Carter 48
40 Reagan 96
41 Bush 48
42 Clinton 96
43 Bush 96

Extracted from *http://www.amstat.org/publications/jse/v13n1/datasets.hayden.html*

Kernel Density Estimates (KDE)

- **Another example**: consider the following data set

Id - USA President – Number of days in office

1 Washington 94

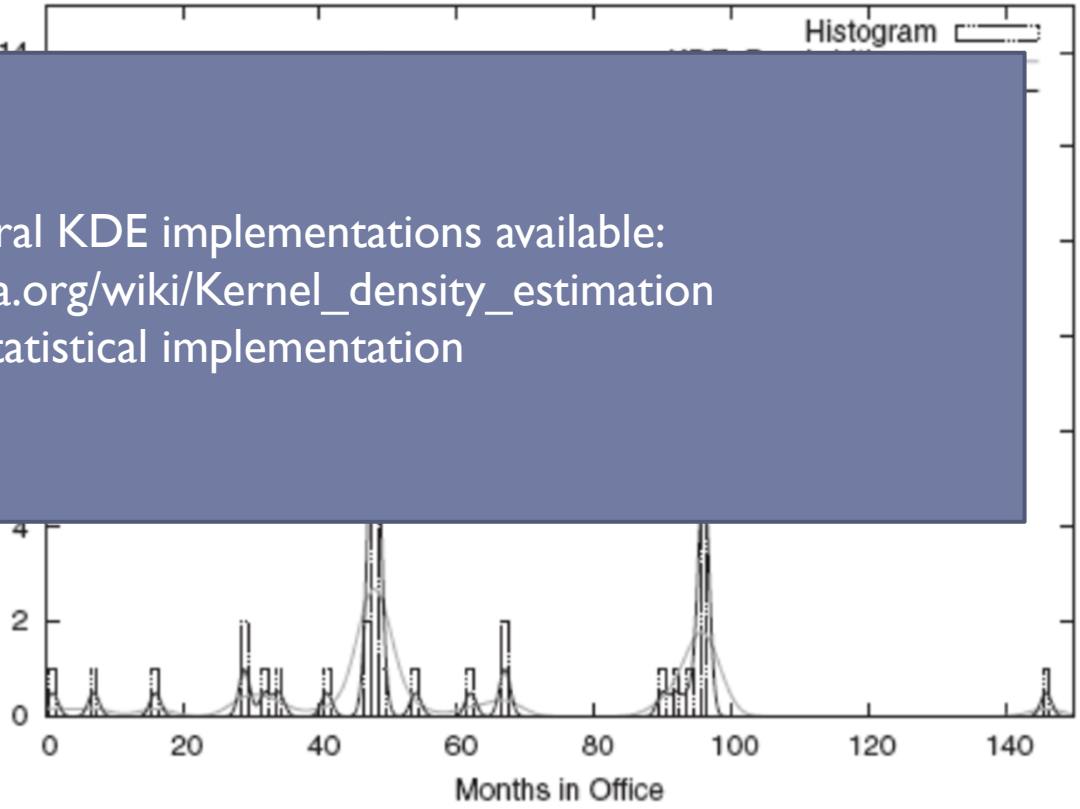
40 Reagan 96

41 Bush 48

42 Clinton 96

43 Bush 96

There are several KDE implementations available:
http://en.wikipedia.org/wiki/Kernel_density_estimation
→ Statistical implementation



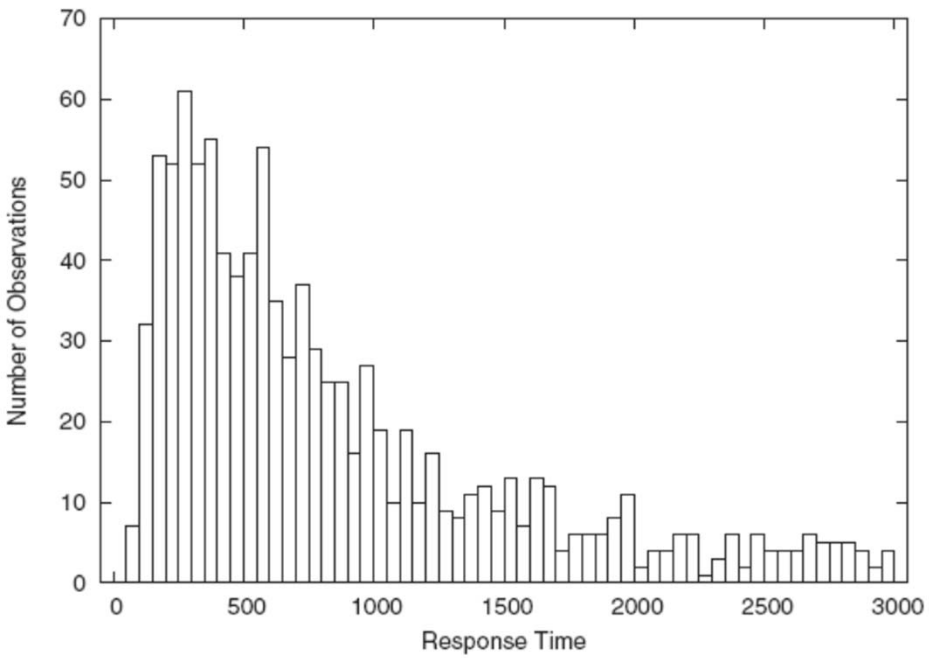
Extracted from *http://www.amstat.org/publications/jse/v13n1/datasets.hayden.html*

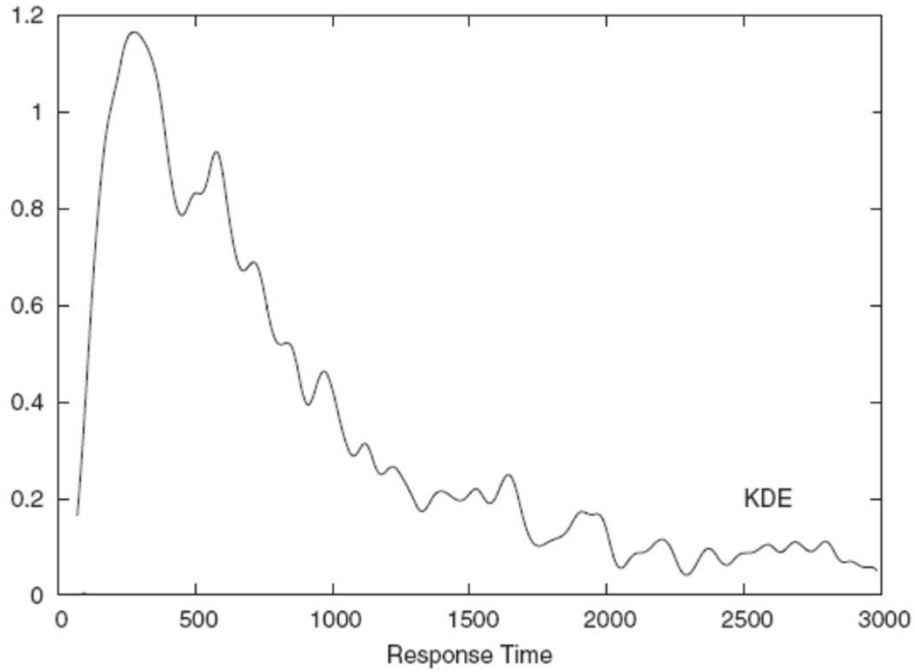
Cumulative Distribution Function (CDF)

- Both histograms and KDE's provide a **good estimate of how probable it is to find a data point with a certain value**
- They provide only estimates, but **knowing exactly the probability of some value range may be desired for proper analysis**; to this end, one can count on Cumulative Distribution Functions
- **For example**, considering the web server example, one might ask: what fraction of requests completes between 150 and 350ms? or under 500ms?
- Relying only on histograms and KDEs, it is necessary to have a good perception of area, a skill which is not well-developed for human beings

Kernel Density Estimates (KDE)

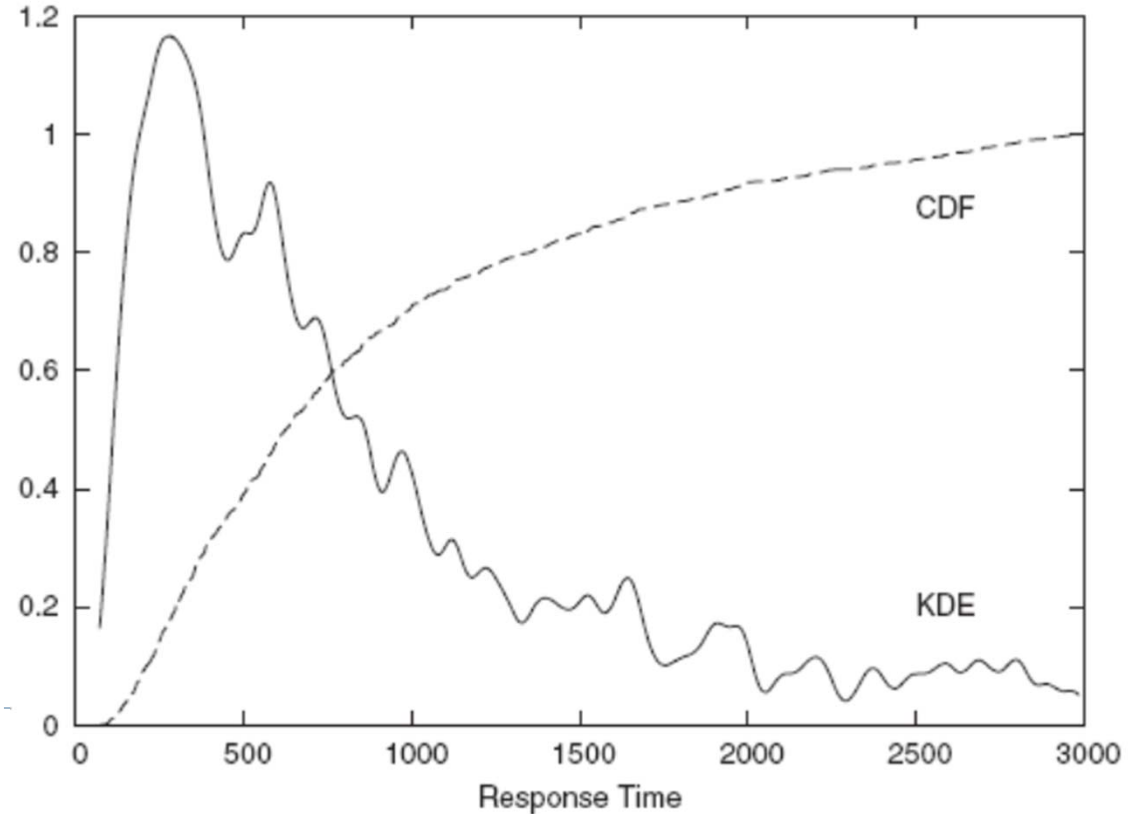
- **Example**: the web server time response data (slide 6) now presented with KDE





Cumulative Distribution Function (CDF)

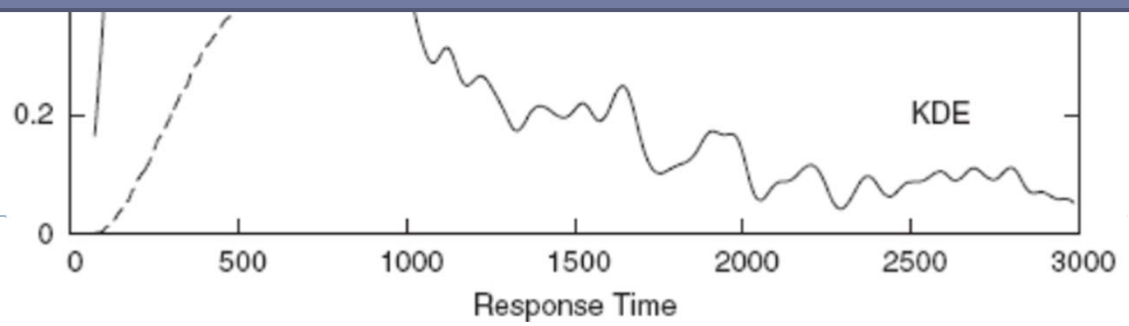
- CDF aids on those problems; given a point x, it tells what fraction of events has occurred "to the left" of x, that is, the fraction of all points x_i with $x_i \leq x$
- For the web server example, we get:



Cumulative Distribution Function (CDF)

➔In the plot, we can see that for t=1,500, the CDF is ~0.85, that is, 15% of all requests take longer than 1,500ms, ~40% take less than 500ms and less than a third are completed in the typical 150-500ms. range

➔These results clarify what is shown in the KDE plot; there it seems that most of the events occur within the peak of t=300 and that the tail so forth contributes little to the answering time – CDF shows that it is the other way around



CDF – Properties

- CDFs are always **monotonically increasing** with x
- They have the **same information** as histogram and KDEs, but in a **different format**
- Different from the histogram, they **do not lose information**
- They approach to zero as x approaches negative infinity, and approach to 1 (100%) as x approaches positive infinity
- CDFs are **unique** for a given data set, different from histograms and KDEs which depend on parameters for their presentation
- They provide information about **how unbalanced is the data set** (if there are, and what are the impacts of the tails)
- CDFs are also known as *lift charts*

CDF – Properties

- **CDFs and histograms are related** according to:

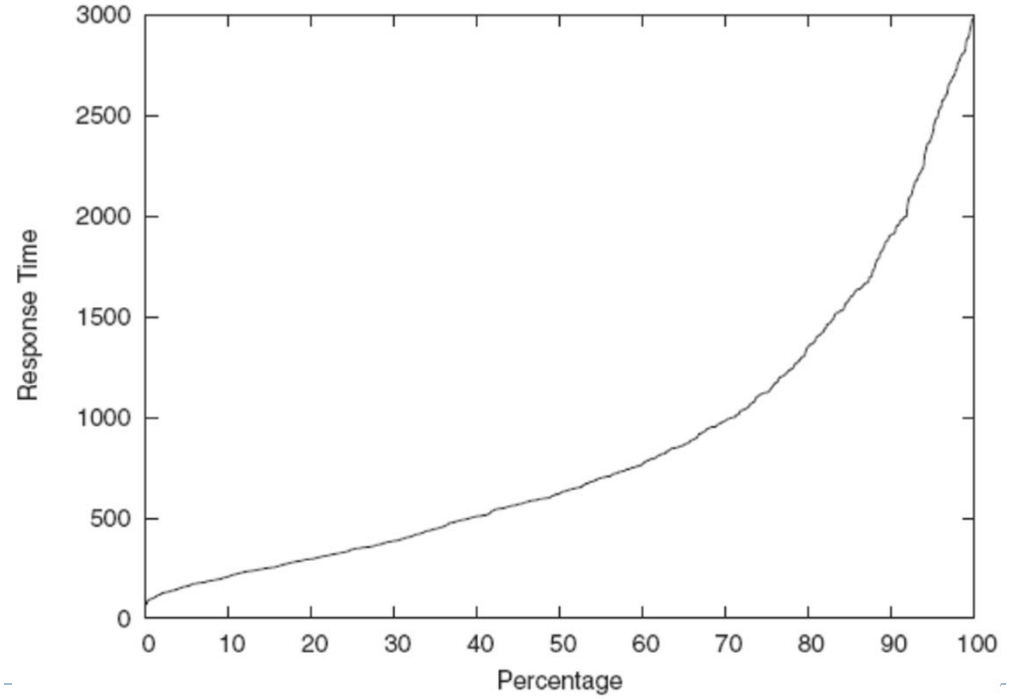
$$\text{cdf}(x) \approx \int_{-\infty}^{x} \mathrm{d}t \, \text{histo}(t)$$

$$\text{histo}(x) \approx \frac{\mathrm{d}}{\mathrm{d}x} \text{cdf}(x)$$

- CDFs are powerful for **comparing the distribution of multiple data sets**, a task that is difficult by other means

CDF – Properties

- By switching the axes of the CDF plot, we get the so-called "**quantile plot**", from which one can easily answer questions such as "What response time corresponds to the 50th percentile?"

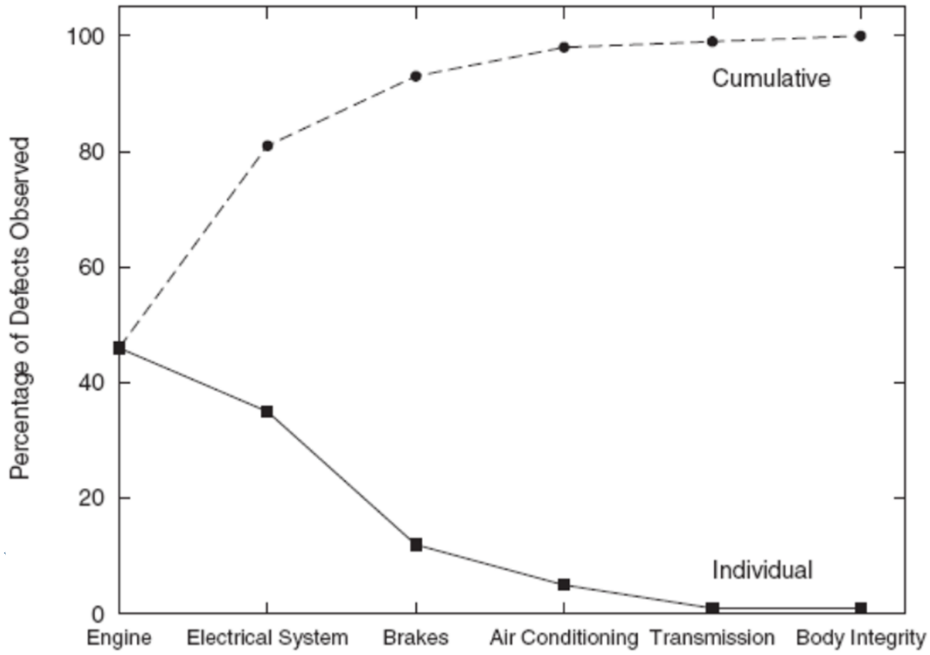


Rank-order plot (Pareto chart)

- Consider the following data: given a set of text manuscripts, **it is possible to count the occurrences of each word**, producing data that is similar to that of the web server example
- However, differently from the webserver data, in which the x axis was defined and ordered according to the response times, **the x axis of textual data makes no sense** if we consider the lexicographical order of words
- In such cases, the solution is to use **rank-order plots**, which are basically the same as histograms and KDE plots, with the difference that the ordering of the x axis **must be given by the counting of the occurrences of each x value**, and not by the values of x
- Terminologically, we say that x is the **independent variable**, while the number of occurrences is the **dependent variable**

Rank-order plot (Pareto chart)

- **For example**, consider a data set that, for each part of a **car system**, it provides the number of defects observed after manufacture
- The plot bellow is a **rank-order-style histogram** together with the CDF plot – it reveals a **pareto-like distribution** according to which only three items answer for over 85% of the events – that is **the company should concentrate their efforts on these elements for a short-term improvement of quality**



References

- Philipp K. Janert, *Data Analysis with Open Source Tools*, O'Reilly, 2010.
- Wikipedia, http://en.wikipedia.org
- Wolfram MathWorld, http://mathworld.wolfram.com/