Aalto University
School of Electrical Engineering

Network Traffic Measurements and Analysis

Lecture I: Data analysis

Markku Liinaharja (slides originally made by Esa Hyytiä)

Department of Communications and Networking
Aalto University, School of Electrical Engineering

Version 0.2, September 20, 2017

Contents

“Abstract” of the lecture

- Measurements have provided a set of numbers - what can we do with those?
- Idea: Basic statistical methods even without much mathematics are sufficient to very many tasks that arise from network measurements

Reality, however,

- Measurement data available is for certain purpose ...
- ...while the current objective something different (cf. traffic classification from flow data)

Objective

- On the one hand, the goal is to learn to utilize basic statistical tools to distill the essential features of measurement data
- On the other hand, the goal is to learn to interpret statistical summaries of measurement data
  - Especially important is to understand the shortcomings of different types of summaries

Objective

- Computational tools
- Network measurements frequently produce **vast amounts of data**:
  - Packet traces without the payload
  - Flow data
- There are many software tools that can be utilized to
  - Manipulate data into a form that is easy to analyze (e.g. scripts)
  - Perform statistical analyses
  - Visualize the results (e.g., gnuplot)
- On this lecture we will use the R software environment for demonstration purposes
  - Downloadable freely: `http://www.r-project.org/`
  - In Ubuntu Linux: `sudo apt-get install r-recommended`

Data preprocessing

- In the network measurement context the numbers are generally obtained by **preprocessing** the raw measurement data
- Preprocessing may include
  - *Cleaning*, e.g., removing incomplete entries
  - *Integration*, e.g., multipoint measurements
  - *Transformation*, e.g., aggregation
  - *Reduction*, e.g., categorization
- Not necessarily difficult, but often more time consuming than the statistical analysis itself!

Contents

Data

- Goal of statistics is to gain understanding from **data**
- Data are numbers with a **context**
  - Statistical tools can be utilized to organize, display and summarize the numbers
  - Understanding of the context is then utilized to draw conclusions of the data

Why network traffic measurements?

- Input for system design (e.g., WWW caches, CDNs, etc.)
- Performance evaluation afterwards
- Anomaly detection, protection against DDoS attacks
- Billing, etc.

Measurement data basics

- Measurement data consists of
  - **individuals** (packet, flow, user,...) and a set of
  - **variables** (packet length, flow size, ...)
- Variables can be
  - **categorical** (protocol type, port number) or
  - **quantitative** (packet size, delay)
- **Distribution** of a variable defines the values the variable can take and how often the variable takes these values
- **Spreadsheet** is a format where each row is an individual and each column is a variable.

Exploratory data analysis (EDA)

- Our approach to measurement data can be considered to be an exploratory one
- EDA: “Describe what you observe”
  - Uncover underlying structure
  - Extract important variables
  - Detect outliers and anomalies
  - Develop (parsimonious) models

Exploring the measurements

EDA approach

- Not to make any assumptions on the data but try to find out what kind of assumptions could be made

Problem $\Rightarrow$ *Data* $\Rightarrow$ *Analysis* $\Rightarrow$ *Model* $\Rightarrow$ *Conclusions*

Other common approaches

- "Classical" statistical approach

Problem $\Rightarrow$ *Data* $\Rightarrow$ *Model* $\Rightarrow$ *Analysis* $\Rightarrow$ *Conclusions*

- Engineering approach:

  Specifications on what and how to measure (e.g., ITU-T)

Basic methodology

- Pre-process the measurements to obtain a spreadsheet
- Rules of thumb (Moore&McCabe):
  - Begin by examining each variable by itself.
    - Then move on to study the relationships of the variables
  - Begin with a graph or graphs.
    - Then add numerical summaries of specific aspects of data

Contents

Single variable analysis

- Begin with a graph or graphs and decide which numerical summaries are the most suitable
- Available tools depend on whether the variable is categorical or continuous

Analyzing a categorical variable

- Categorical variables have two qualities to analyze:

  Counts: How many instances there are in each category

  Percents: What are the relative shares of instances in each category

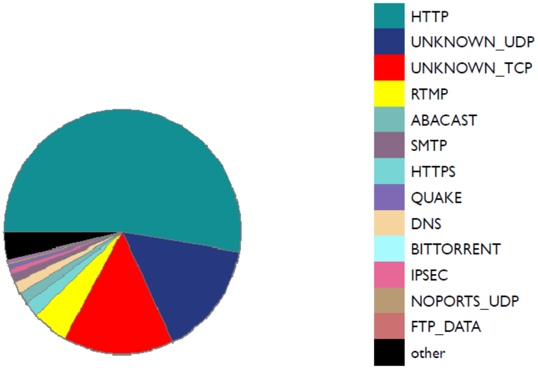
- Generally with low number of instances the counts are more interesting and with large number of instances the percents are more informative
  - Exceptions include e.g. cases where we are interested only in one category

Visualizing a categorical variable

- Categorical variables are easy to grasp from a bar graph or from a pie chart
- Rules of Thumb:
  - Use bar graphs if the actual number of instances is relevant
  - Use pie charts if the proportions are more interesting.
  - If the categories do not have a "natural order", it is often convenient to visualize the data so that the categories are ordered according to their relative frequency

Example

- Pie chart of application packets in Chicago monitor A (www.caida.org) on March 10 - March 11 2011



Analyzing a quantitative variable

- Quantitative variable is analyzed by the distribution
- Always try to plot your data first

  1. Look for overall pattern
  2. Look for deviations from the pattern
  3. Produce a numerical summary to briefly describe center and spread of data

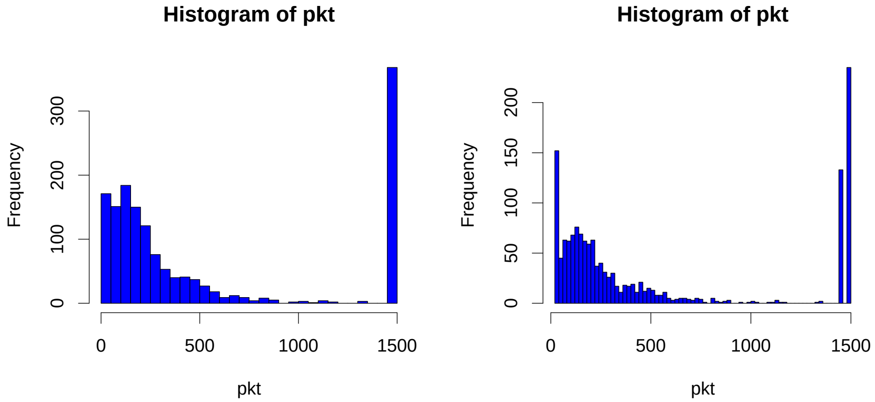
- Simplest approach: Raw data plot
- Plot values "one-by-one"
  - Discrete or continuous?
  - Range, spread?
  - Special values?
- Is not a meaningful description of the distribution by itself

Visualization: Histogram

- Describes the distribution of the variable
- Break the range of values of a variable into classes and count the measurements that fall into each class - frequency table
- Plot the frequencies (or normalized frequencies) using bars
- Use your own judgment in choosing the classes (a.k.a. bins)
- Goal: the distribution should be well illustrated
- The appearance of the histogram may change significantly you change the classes
- Try different selections

Example

```
> pkt<-scan("packetsizes.txt")
> hist(pkt,col=4,breaks=50)
```

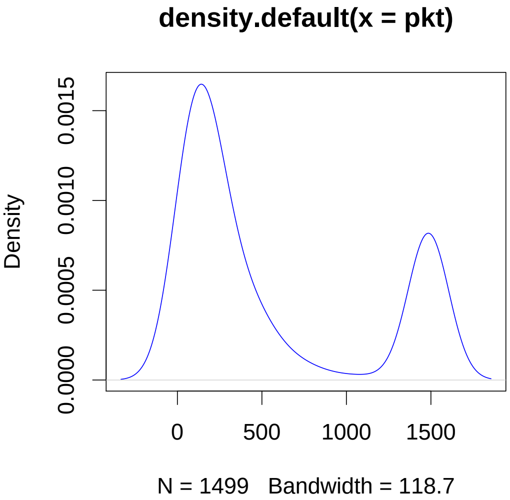


Visualization: Density curves

- Histogram depends heavily on the class selection
- Another useful tool for characterizing the data is to estimate the density curve of the underlying variable
- Good for “smooth” distributions
- Density curve has an area of 1 (pdf!)
- Always non-negative
- Determined by statistical softwares such as R

Example (of not so good density curve)

```
> plot(density(pkt), col=4)
```



Examining the distribution: Pattern

- Overall pattern
- Shape
  - One or several modes
  - Symmetricity, skewness
- Center
  - Where the distribution lies, “mean value”, “typical value”
- Spread
  - How much the values vary

Outliers, outlying observations

Outlier is an observation that is **clearly** outside the overall pattern

- Outlier can result from a measurement error ... or not
- Outliers can significantly complicate the numerical description and analysis of data
  - Screen the data and remove outliers (careful!)
  - Use robust statistics to describe the data

Describing the distribution with numbers

- Distribution shape is described by inspecting the histogram
- Numbers are generally used for the center and spread
- Remember: The numbers and graphs are aids to understanding the data not the goals themselves

Measuring center

- For measurements $\{x_i\}$, the center can be described by the **sample mean**

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

- Sample mean is an optimal estimator of the center of the normal distribution
- However, the mean is sensitive to the influence of outliers
  - Even one gross error can make the mean arbitrarily large
  - The **breakdown point**[1] of mean is 0%

[1] The breakdown point is the fraction of incorrect (arbitrarily large) observations an estimator can handle before going haywire.

Measuring center

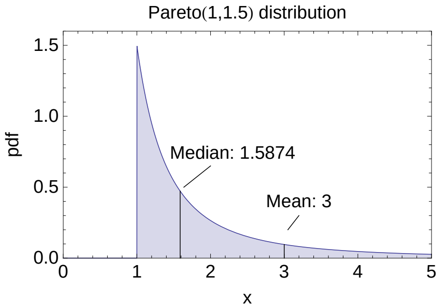
- Robust methods provide alternative way of characterizing the center
- **Median** is defined as

  the centermost value of the ordered data.

  If the number of data is even, the median is the mean of two centermost values.
- Median is more resistant measure of the center
  - It has breakdown point of 50% (it can tolerate 50% of gross errors before becoming arbitrarily large)

Mean vs. Median

- Mean is the "average" value of the variable whereas the median is the "typical" value
- If the distribution is symmetric, both are close to each other (and identical if the distribution is exactly symmetric)
- If the distribution is skewed, mean tends to be farther out in the long tail



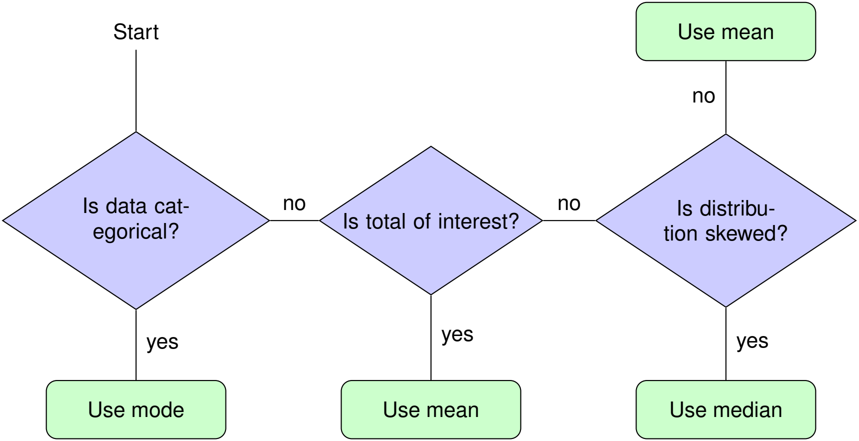


Figure: Selecting measure for center.

Note #1

Suppose you have managed to improve transfer rate in a communication system:

Wireless layer (MAC)	3%
Network layer (IP)	10%
Application layer	50%

Q: What is the average improvement?

- $(3\% + 10\% + 50\%)/3 = 21\%$?
- No, improvements are multiplicative!
- $(1.03 \cdot 1.10 \cdot 1.50)^{1/3} \approx 1.193 \Rightarrow \underline{19.3\%}$

This is the so-called **geometric mean**,

$$(a_1 \cdot a_2 \cdot \ldots \cdot a_n)^{1/n}.$$

Note #2

Suppose you have measured a transfer time of a file n times:

file size	b [byte]
transfer time	t_i [sec]
mean rate	$r_i = b/t_i$

Q: What is the average transfer rate?

- Mean of sample rates?

$$\frac{r_1 + \ldots + r_n}{n} = \frac{b/t_1 + \ldots + b/t_n}{n}$$

- No, a better metric is the **harmonic mean**

$$\frac{n}{1/r_1 + \ldots + 1/r_n} = \frac{nb}{t_1 + \ldots t_n}$$

Measuring spread

- Simplest useful numerical description of data consists of both a measure of center and a measure of spread
- Easiest way of describing spread is to give several percentiles. The *p*:th **percentile** is the value such that p% of the measurements fall at or below it
- Median is the 50th, first **quartile** (Q1) is the 25th and third quartile (Q3) is the 75th percentile
- **IQR** = Interquartile range, Q3-Q1

The five number summary and boxplots

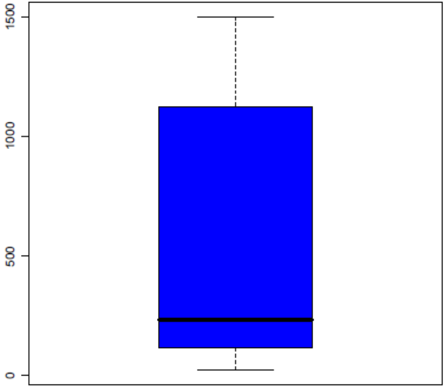
- The **five number summary**,

  (Minimum, Q1, Median, Q3, Maximum)

  is a good summary of the distribution as a whole
- The five number summary is generally depicted using boxplots:
  - Central box spans the quartiles Q1 and Q3
  - Line marks the median
  - Lines extend from the box to mark the maximum and minimum
  - Conventionally observations more than 1.5 times the inter-quartile range (Q3-Q1) from the median are plotted separately
  - Especially suitable for comparison of distributions

Example

```
➢summary(pkt)
 Min.    1st Qu.   Median   Mean    3rd Qu.   Max.
 22.0    115.5     233.0    536.9   1123.0    1500.0
➢boxplot(pkt,col=4)
```



Measuring spread: Standard deviation

- **Sample standard deviation** is the most common measure of spread

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- **Sample variance** is given by s^2
- Why squared distances from mean?
  - Sum of the squared deviations from mean smaller than from any other point
  - Optimal measure in normal distributions
- Why dividing by $n-1$ and not by n?
  - Sum of unsquared deviations is zero, so if you know $n-1$ deviations you immediately can derive the missing one
  - There are $n-1$ degrees of freedom
  - Important to remember when n is small

Notes on standard deviation

- Should be used only when using mean as a measure of the center
- Like mean, standard deviation is not robust against outliers
- Squared deviations make the situation even worse!
  - For the packet data distribution on page 21 $s = 569.27$
- Note that variance σ^2 of random variable X is

$$\sigma^2 \triangleq \mathrm{E}((X - \mu)^2),$$

where μ is the mean, $\mu = \mathrm{E}(X)$. Sample variance s^2 is an **unbiased** estimator for σ^2.

Coefficient of variation (c_V, C.O.V.)

Sample mean *m* and standard deviation *s* include a unit (e.g. bytes or seconds)

Coefficient of variation (C.O.V):

Coefficient of variantion is a *dimensionless* measure of spread

$$\boxed{c_V = \frac{\text{standard deviation}}{\text{mean}} = \frac{s}{m}}$$

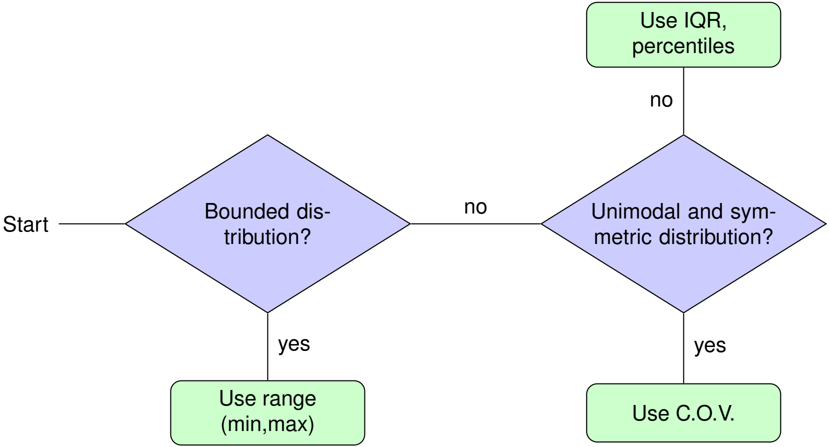


Figure: Selecting a measure for spread.

Linear transformations

- A linear transformation of type

$$x_{\text{new}} = ax + b,$$

  does not change the shape of the distribution
- The transformation has the following effects on the statistics
  - Measure of spread is multiplied by a
  - Measure of center, m, becomes $am + b$
- Useful e.g. in changing the unit of measurement
  - bits vs. bytes
  - packet size vs. payload

Choosing the measures of center and spread

- Rule of thumb: Do not use mean and standard deviation for strongly skewed distributions
- Distributions with multiple peaks or gaps are ill-suited for simple numerical description in general
- Numbers report specific facts about a distribution, a graph is generally more informative than plain numbers when it comes to the overall picture

Examples of other measures

- **Skewness** and **kurtosis**, i.e., the third and fourth standardized moments of a distribution,

$$\boxed{\frac{\mathrm{E}((X-\mu)^k)}{\sigma^k}, \quad \text{where } k = 3, 4.}$$

- Describe the shape of the distribution
  - Skewness < 0, tail to the left
  - Skewness > 0, tail to the right
  - Kurtosis low $\Rightarrow$ flat but short tailed distribution
  - Kurtosis high $\Rightarrow$ sharp with heavy tails

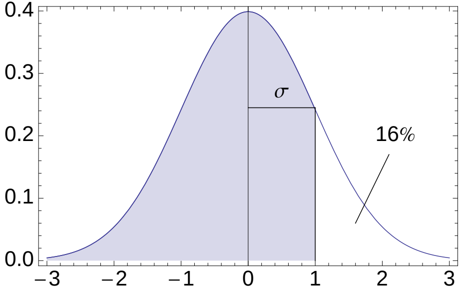
Examples of other measures

- Trimmed mean
  - Mean of the central $1 - 2\alpha$ part of the distribution
  - If outliers cannot be removed one-by-one
  - Uses more information than median
- Median absolute deviation (MAD)

$$\mathrm{MAD} = \mathrm{Median}_i(|X_i - \mathrm{Median}_j(X_j)|).$$

Intuition on center and spread for density curves

- Mean is the "balance point" of the density curve
- Median is the point that divides the area of density curve into two equal parts
- Standard deviation for normal curves: the "turning point"



Empirical cumulative distribution function

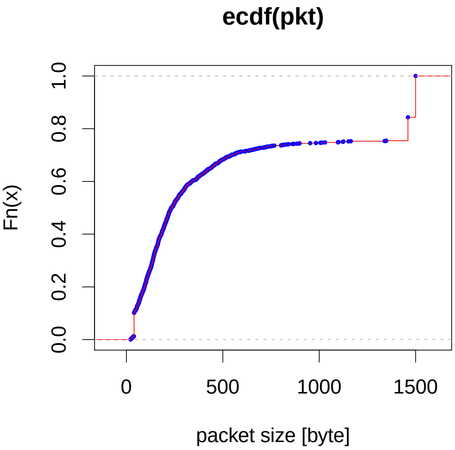
- Another important way of characterizing the distribution of a variable is the empirical cumulative distribution (cdf)

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x).$$

- Here the function $I(x_i \leq x) = 1$ if $x_i \leq x$, otherwise 0

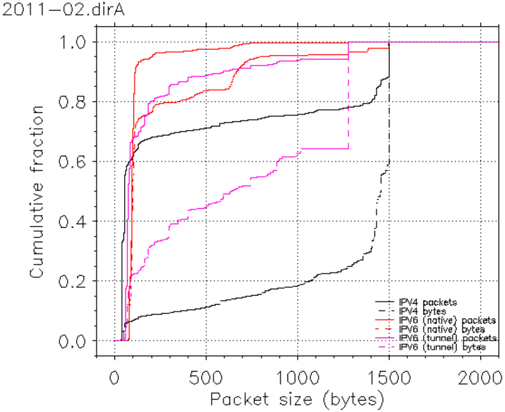
Example

```
> pkt<-scan("packetsizes.txt")
> fn<-ecdf(pkt)
> plot(fn,verticals=TRUE,col.points=4,col.hor=2,col.vert=2,cex=.4)
```



Example (www.caida.org)

- Packet size distribution on a backbone link
- Feb 17 2011
- 12:59:04 - 14:01:04



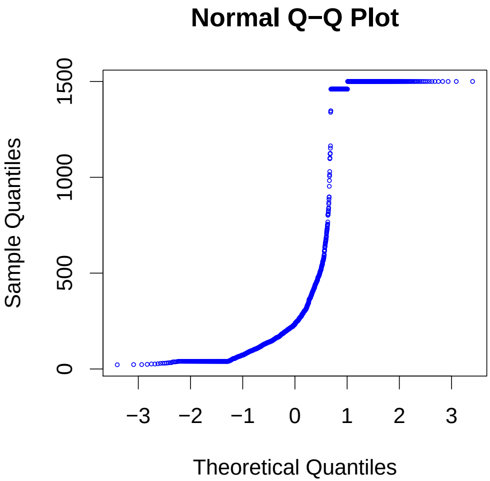
147986351749 bytes
2221625 bytes (1.5×10^{-3} %)
852714 bytes (5.8×10^{-4} %)

Quantile plots

- **Quantile plots** (aka Q-Q plots) are a useful tool in comparing whether your measurements can be described by a certain statistical distribution (or by another data set)
- More about distributions on the next lecture!
- For data of n measurements plot (x, y), where:
  - x: $k/(n+1),\ k = 1, \ldots, n$ quantile of the comparison distribution
  - y: order statistics of data, i.e., k:th smallest measurement
- If the points constitute a straight line, the distributions are "similar" (and if the line is close to the 45 degree line, the distributions are identical)

Example: Normal Q-Q plot

```
> pkt<-scan("packetsizes.txt")
> qqnorm(pkt, col=4,cex=0.4)
```

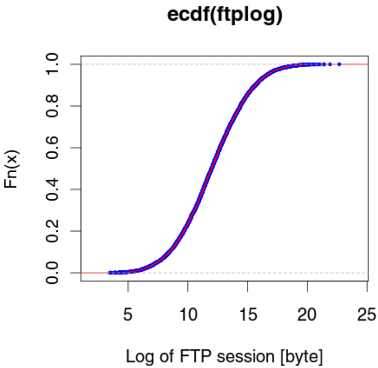


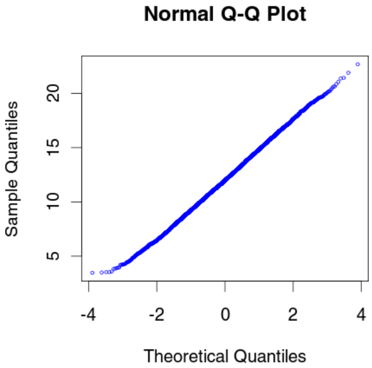
Challenges with measurement data

- Data volume
  - Storage, handling, overplotting, ...
  - Some problems can be avoided by
    - Preprocessing steps (reduction, aggregation, ...)
    - Sampling
- High variability
  - Causes instability for many metrics
    - Use robust statistics
  - Makes distribution plots less illustrative
    - Use logarithms in plots

Example: Q-Q log-normal

```
> ftp<-scan("ftpsessions.txt")
> ftplog<-log(ftp)
> fn<-ecdf(ftplog)
> plot(fn,verticals=TRUE,col.points=4,col.hor=2,col.vert=2,xlab="...")
```

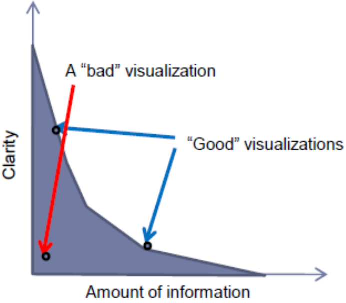




$\Rightarrow$ Log-normal distribution!

In reporting the results ...

- Suitable numerical summaries support understanding but may oversimplify some aspects of the data
- Graphs are efficient in reporting measurement results
- When visualizing the data
  - Make sure that the graph is as informative as possible
  - "Informativeness" sets a appropriate balance between amount of information and clarity

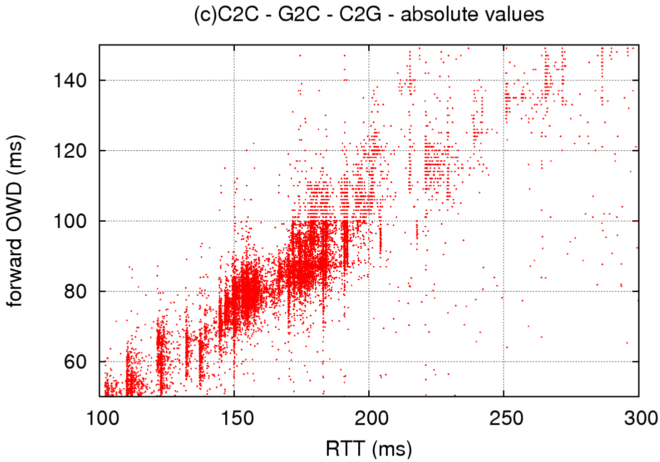


Contents

Relationships between variables

- The most important tool for studying relationships between variables is the **scatterplot**
  - Values of one variable are plotted on the x-axis and values of the other variable are on the y-axis
  - Conventionally x is the *explanatory variable* and y is the *response variable*
- Categorical variables can be added using different markers or colors
  - Sometimes referred to as *Youden plot*

Example scatterplot



Source: A. Pathak, et al., *A Measurement Study of Internet Delay Asymmetry*, PAM 2008. [use plot() in R or gnuplot to produce scatterplots]

Analysing a scatterplot

- Form, direction and strength of the relationship
- Form
  - Linear, curved,...
  - Clusters
- Direction
  - Positive or negative association?
- Strength, outliers
  - How close the points are to the form?

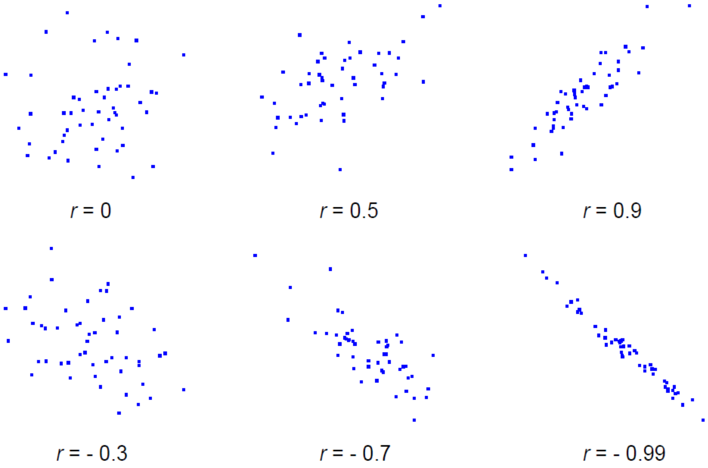
Correlation

- Correlation between two variables x and y,

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right).$$

- Gets values between $[-1, 1]$
  - Values near 0 indicate weak linear relationship
- Describes only linear relationship
- Not resistant
- r^2 is the fraction of variation in data that is explained by least-squares regression of y on x, r is the slope

Example: correlation



Caution on relationship analysis

- Outliers and influential observations
- Correlations on averages usually higher than on individuals
- Lurking variables
- High correlation does not imply **causation.** Possible associations
  - Causation
  - Common response
  - Confounding
- Establishing causation
  - Conduct an **experiment**, where the effects of lurking variables are controlled

Contents

Multi-dimensional data

- Measurement data usually contains many variables for each individual
  - Try to reduce the number of variables in preprocessing
- In certain cases we need to preserve the information $\rightarrow$ multi-dimensional data
  - No prior knowledge what to look for
  - The studied phenomena span over several variables, e.g., intrusion detection

Exploring multi-dimensional data

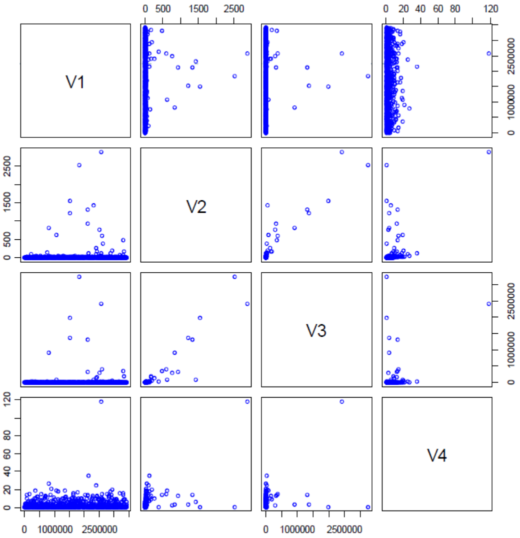
- The general rule applies: **Plot your data!**
- The plots are typically not as intuitive as with one and two variables, visualization of multi-dimensional data is a challenge
- Methods:
  - Multi-dimensional plots
  - Projection methods

Example: Pairs plot

Matrix of scatterplots

Variables:

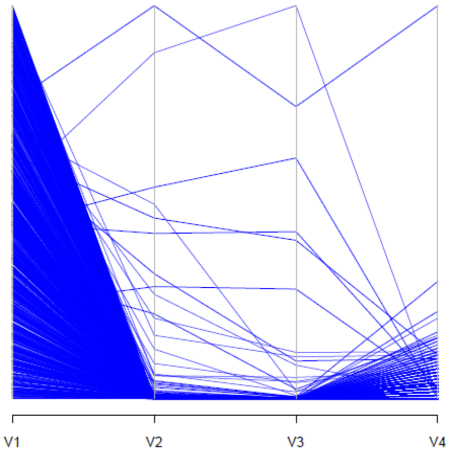
1. Source ID
2. Packets
3. Bytes
4. Flows



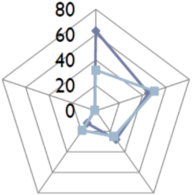

```
➤pairs(flowdata,col=4)
```

Parallel plots

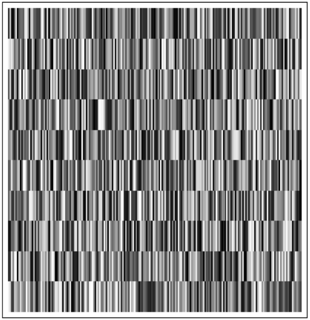
- Each measurement is represented by a horizontal path



Other plots



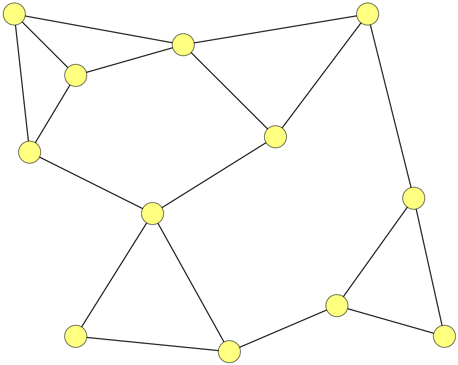
Radar / star plots



Color histograms

Multi-dimensional scaling

- A method to visualize multi-dimensional proximity data in low dimension
- E.g., *ad hoc* network MAC neighbors topology

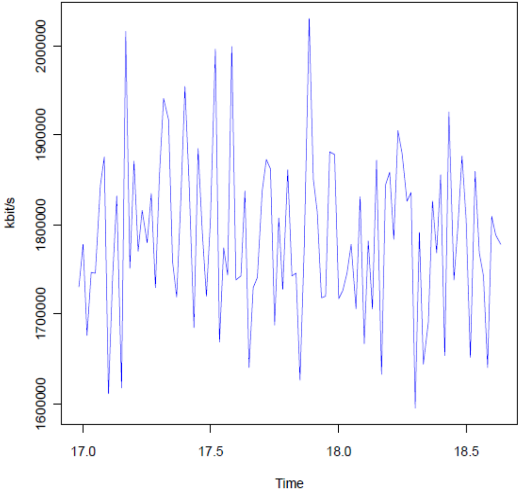


Contents

Time plots

- Whenever data are collected over time it is a good idea to plot the measurements in time order
- Distribution studies ignore the time order, which may be misleading when there is a systematic change in time
- For example, if traffic load is high at a certain moment, it is likely that it is still high a second afterwards

Example: Time plot



Time series analysis

- Analysis of data ordered by the time the data were collected
- Usually equally spaced time instants (discrete time)
- Goals:
  - Modeling: To determine the process that has produced the data
  - (Forecasting: Point estimates and confidence intervals)
- Exploratory aspects
  - **Memory** and **stability** of the data

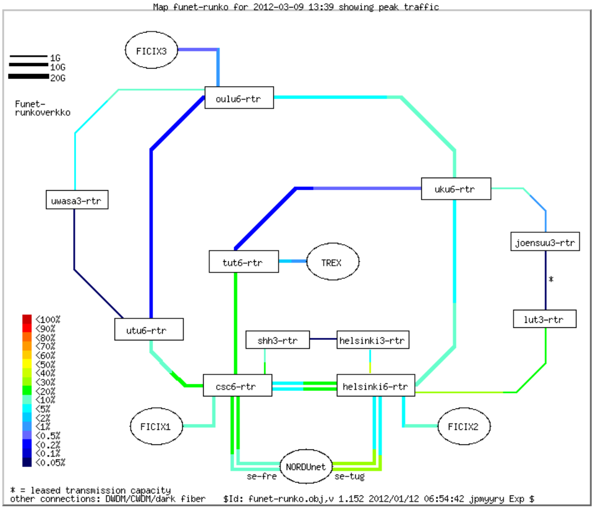
Stability of data

- Stability refers to the traffic consistency over time
  - Mean, variance, etc. do not change over time
- Distinct from **stationary**, which is a formal property of a stochastic process
  - If the data are stable, a stationary model may be applicable
- Subjective concept
  - Can be tested roughly by dividing the data into successive batches and analyzing whether some parameter estimates remain roughly constant in all the batches

Traffic at long time scales

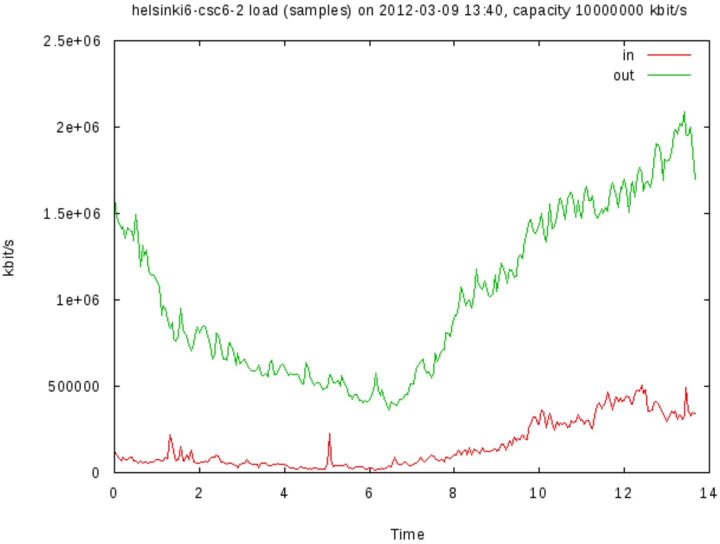
- Network traffic is not usually stable at long time scales
  - Long time scale trend of “ever increasing traffic”
  - Predictable components
  - Daily cycle
  - Weekly cycle
  - Yearly cycle (summer holidays etc.)
  - Special events (football world championships etc.)
- Unpredictable **external** effects
  - Accidents
  - Network failures

Example: Funet backbone network



http://www.csc.fi/funet/status/tools/wm

Time plot example: csc6rtr – helsinki6rtr



Decomposition of time series

- Statistical tools can be utilized to decompose the series into components
- Trend
- Seasonal variation
- ...
- Irregular influences

Traffic at short time scales

- At short time scales, comparable to minutes, traffic can be usually described as a stationary stochastic process
- However, networks contain buffers and control algorithms that maintain past history in a way it affects the current behavior
- Short-range memory and long-range memory are often both present at network measurements

Memory in system behavior

- Memory has both good and bad effects
  - Good: Near future more predictable
  - Bad: The amount of information in each measurement decreases, high variability
  - …the Ugly?

Analyzing memory: Lag plots

- The easiest way to observe short-range memory is to consider **lag plots**
- Plot X_k against, e.g., X_{k+1}
- Check randomness, outliers, deterministic models

With R:

```
> lag.plot(x1)
```

Analyzing memory: ACF

- **Empirical autocorrelation function** (ACF) is defined as

$$\hat{r}(k) \triangleq \frac{\frac{1}{N-k}\sum_{i=1}^{N-k}(x_i-\bar{x})(x_{i+k}-\bar{x})}{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}.$$

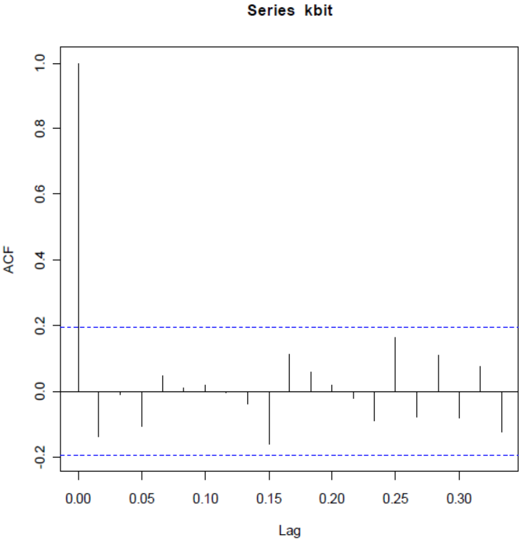
- Estimate for the normalized autocovariance function,

$$r(k) \triangleq \mathrm{E}\left(\frac{(X_i-\mu)(X_{i+k}-\mu)}{\sigma^2}\right).$$

With R:

```
> plot(acf(x1))
```

Example: lag



Series kbit
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Lag

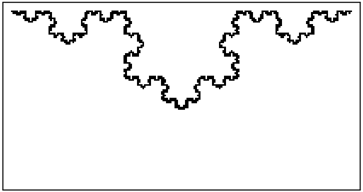
Examining the ACF plot

ACF plot can be used to assess

- Are the data random?
- Are the adjacent measurements related?
- What model could be appropriate?
- Are the data self-similar?
  - Is ACF similar at different time scales?

Self-similarity

- Network traffic is often self-similar
  - Its statistical properties remain same under “zooming”
  - Cf. Koch curves, ferns, coast lines etc.
- Results essentially from long-range dependence
- We will return to self-similarity in more detail later as stochastic processes are discussed



Koch curve (Source: Wikipedia)

Literature

- David S. Moore and George P. McCabe, Introduction to the practice of statistics, 5th Edition, W.H. Freeman & Co., 2006
  - Chapters 1-2
- NIST/SEMATECH, Engineering Statistics Handbook, Chapter 1, `http://www.itl.nist.gov/div898/handbook/index.htm`