Aalto University
School of Electrical
Engineering

Network Traffic Measurements and Analysis

Lecture II: Sampling and experimental design

Markku Liinaharja (slides originally made by Esa Hyytiä)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

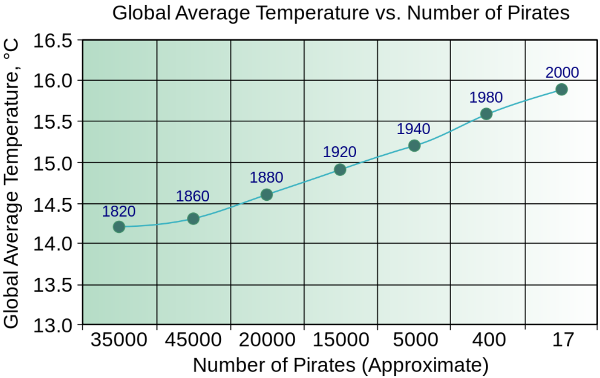
Experiments and observational studies

- **Experiments** differ from **observational studies** by the active imposition of some treatment on the subject of the experiment
- In this course, we will discuss the design of experiments only briefly and concentrate more on sampling

Question of causation

- In observational studies we may study the **association** of variables but it may be difficult to establish **causation**
  - **Common response**: a lurking variable causes the association between the observed variables
  - **Confounding** of two variables means that we cannot distinguish their effects on the response variable
  - Even a very strong association is not, by itself, good evidence of a cause-and-effect link
- Causation can be established by experiments
- If experiments are not practically feasible, causation requires very good motivation
  - Strong consistent association in many studies, clear explanation of the alleged causal link, careful examination of alternatives

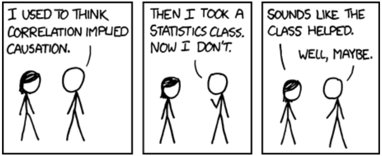
Correlation does not mean causation (1)



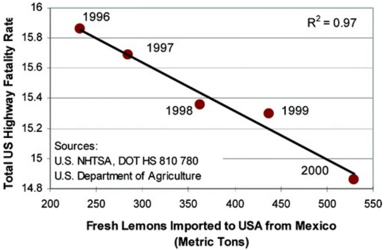
(Source: wikipedia.)

- The number of pirates causes the global warming?

Correlation does not mean causation (2)



Source: `http://xkcd.com/552/`



Source: J. Chem. Inf. Model., 2008, 48 (1), pp 25–26
`http://pubs.acs.org/doi/abs/10.1021/ci700332k`

Chocolate Consumption vs. Nobel Laureates

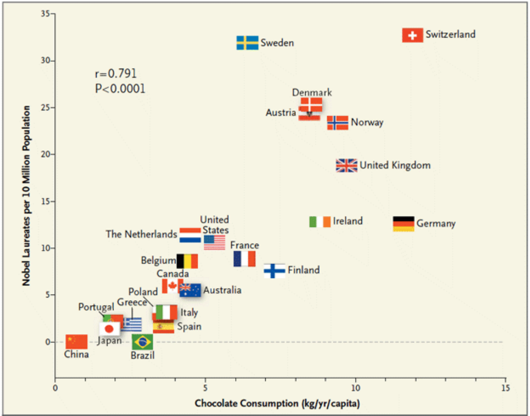


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Source: The New England Journal of Medicine

`http://www.nejm.org/doi/full/10.1056/NEJMon1211064`

Contents

Experimental design

- Reveal the response of a variable (response variable) to the changes in other variables (factors or explanatory variables)
- In an experiment one or more treatments are imposed on **experimental units** (subjects)
  - A treatment is a combination of levels of variables (often called factors)
- The advantage of experiments over observational studies is that we can focus on the specific factors we are interested in while the effects of lurking variables can be controlled

Example

- TCP Vegas has a better performance than TCP Reno, why?
- Response variable: **Throughput**
- Factors / explanatory variables
  - New retransmission mechanism, congestion recovery
  - Congestion avoidance mechanism
  - Modified slow-start mechanism
- Levels of factors
  - ON-OFF
- Treatment
  - E.g., congestion avoidance + modified slow start
- Experimental units
  - Simulation runs/measurements

Ref.: Hengartner, U., Bolliger, J., Gross, T.: TCP Vegas revisited. In IEEE INFOCOM'00 (2000).

Principles of experimental design

- **Control** the effect of lurking variables
- **Randomize** the subjects into the treatments
- **Repeat** each treatment on many units to reduce chance variation in results

Experimental designs: Control

- Comparison is the simplest form of control
  - Compare two or more treatments in the same environment
  - A zero-treatment group is called a **control group**
    - The same lurking variables operate also on the control group
- Factorial design
  - Evaluate all possible combinations of factors
  - Effects of each factor and interactions
  - E.g., TCP Vegas vs. TCP Reno
    - Three ON-OFF factors
    - $2^3 - 1$ different treatments

Experimental designs: Randomization

- Completely randomized
  - Units are allocated at random among all treatments
- Matched pairs (two treatments)
  - Find matching pairs of units based on some facts and randomize the treatments within the pairs
- Generalization to larger groups is called block design
  - “Blocking” tries to eliminate sources of variability that are not of interest in the experiment

Experimental designs: Repeat

- Increasing the number of experimental units in each group will reduce the probability that the differences in response variables occur by chance
  - The effects of chance will average out
- The more subtle the actual difference the more units are needed to recognize the difference

Cautions about experimentation

- Careful attention to the detail
  - An experiment can influence the outcomes in unexpected ways
- Lack of realism
  - An experimental setup may not really duplicate the conditions where we want to apply the results
  - Statistical analysis of an experiment does not tell how the results can be generalized

Contents

Sampling

- Sampling selects a **sample**, a part of the **population** of interest to represent the whole
- Very useful when time, cost, or inconvenience of analyzing the whole population is prohibitive
  - In network measurements there is usually plenty of data:
    - Sampling may be utilized to facilitate computations, reduce the storage requirements or help online measurements on high-speed links
  - When aiming at general results, all measurements are just samples!
- Poorly designed sampling gives misleading conclusions on the population
- Several basic sampling designs; simple random sampling, stratified sampling, deterministic sampling

Simple random sampling

- **Simple random sample (SRS)** of size n consists of n individuals chosen from the population in such a way that every set of n individuals has equal chance to be the selected sample
  - Each treatment group in completely randomized design is an SRS
- Generalization: Probability sample
  - Impersonal chance: Each individual has a certain probability to become selected in the sample
  - In SRS the probabilities are equal

Stratified sampling

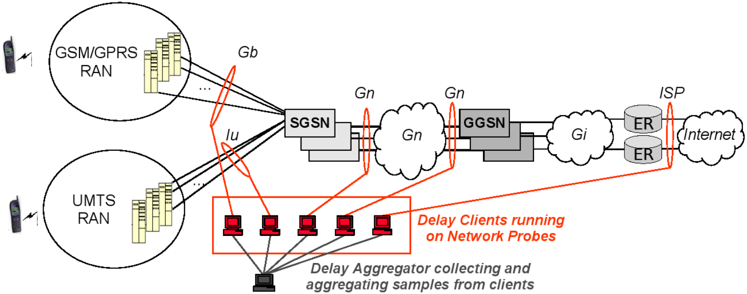
- **Stratified random sample** is selected by first dividing the population into groups of “similar” individuals, called strata. Then SRS is selected from each stratum and combined to form the full sample
- Strata are chosen based on some external facts that are known before the sample is taken
  - Stratified sampling provides more accurate results than SRS by utilizing the external knowledge
- Generalization to successive group selection: Multistage samples
  - Each stage narrows down the population by a sampling method
  - Produces a sample that is a cluster of individuals

Deterministic sampling

- **Deterministic sampling** selects individuals based on certain deterministic rules
  - E.g., select every *n*:th packet traversing through the measurement point
  - Simple to implement, but may introduce bias, systematic error in a way sample represents the population
- Variation: Systematic random sampling
  - Systematic random sampling selects random starting point and continues from there using a deterministic rule
  - Guarantees coverage, but weak against periodicities

Example: One-way delay measurement setup

A **multipoint measurement** setup, where *packetID* (via a hash function) is utilized. Packet sampling based on *packetID* is optional.



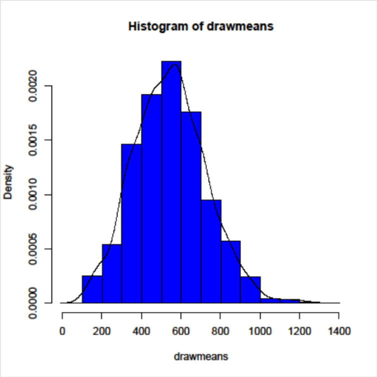
From *P. Romirer-Maierhofer and F. Ricciato, Towards Anomaly Detection in One-Way Delay Measurements for 3G Mobile Networks: A Preliminary Study, IP Operations and Management, LNCS, 2008, (5275/2008).*

Statistic and Sampling distribution

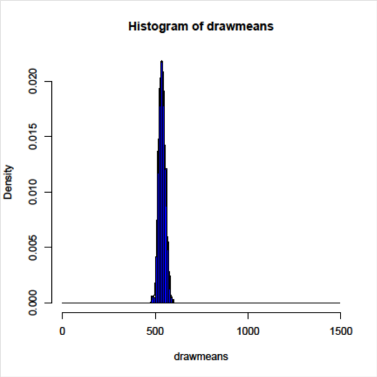
- **Statistic** is a number that describes a sample
  - Used to estimate a parameter of the population
- Sampling variability
  - Value of a statistic varies from sample to sample
- **Sampling distribution** of a statistic is the distribution of the values over all possible samples of same size

Example: Sampling distribution

1000 samples of size 10



1000 samples of size 1000



Bias and variability

- **Bias** concerns the center of the sampling distribution
  - Statistic used to estimate a parameter is unbiased if the mean of sampling distribution is equal to the true value of the parameter
  - Can be reduced by using random sampling
- **Variability of a statistic** is described by the spread of the sampling distribution
  - Can be reduced by using larger samples

Sampling distribution for counts and proportions

- Consider a large population
- We make observations each of which falls into one of two categories:
  - “success” or
  - ”failure”
- Probability for ”success” in population is p
- Estimate p!

Distribution of counts

- Consider the "success" count X in samples of n independent observations
- Number of "successes" has (given a large population, at least 20 times the sample) the distribution $\mathrm{Bin}(n, p)$,

$$p_i = \mathrm{P}\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}.$$

Counts mean	np
Standard deviation	$\sqrt{np(1-p)}$

Counts and proportions

- Consider the **sample proportion** statistic

$$\hat{p} \triangleq \frac{\text{"successes"}}{\text{"samples"}} = \frac{X}{n}.$$

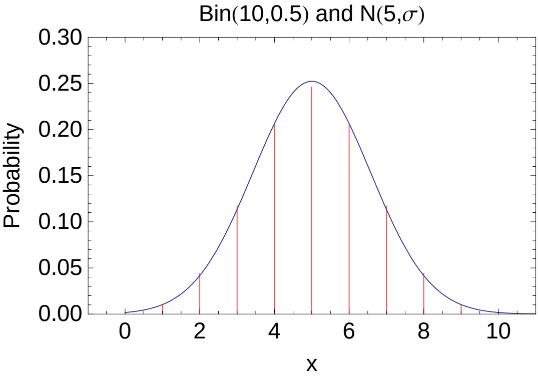
- We have:

Proportion mean	p
Standard deviation	$\sqrt{\frac{p(1-p)}{n}}$

- Sample proportion
  - Unbiased estimator for p
  - Spread goes to 0 as sample size grows – large enough sample will give an accurate estimate of p

Sampling distribution of proportions

When np (and $n(1-p)$) increases, the sampling distribution of proportions approaches the normal distribution



Sampling distribution of sample mean

- Averages
  - Less variable than individual observations
  - More normal than individual observations
- Sample mean of X_i with mean μ and standard deviation σ

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Mean	μ
Standard deviation	$\frac{\sigma}{\sqrt{n}}$

Central limit theorem (CLT)

- Sampling distribution of sample mean is approximately normal for any population with mean μ, assuming
  - Finite standard deviation σ
  - Large n

$$\bar{x} \simeq \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

- More generally, a distribution of a sum of many small random quantities is close to normal

Sampling and network measurements

- The main advantages of sampling are that it reduces
  - Required amount of data
  - Required measurement effort
- The variability of a statistic does not depend on the size of the (large enough) population!
- Potential pitfalls
  - Failing to produce random samples
    - Sampling processes may introduce bias, e.g. periodic measurements
  - Measurement conditions not representative
    - Data set may not describe the population

Literature

- David S. Moore and George P. McCabe, **Introduction to the practice of statistics**, 5th Edition, W.H. Freeman&Co., 2006, Chapters 3, 5