DM63
HEURISTICS FOR
COMBINATORIAL OPTIMIZATION

Lecture 12

Empirical Analysis

Marco Chiarandini

Outline

1. Descriptive Statistics
   Computer Graphics for Sampled Data
   Correlation and Linear Regression
   Characterization

2. Inferential Statistics
   Two Sample Tests
   Analysis of Variance

Outline

1. Descriptive Statistics
   Computer Graphics for Sampled Data
   Correlation and Linear Regression
   Characterization

2. Inferential Statistics
   Two Sample Tests
   Analysis of Variance

Descriptive Statistics

Samples $X^n = (x_1, \ldots, x_n)$ are used to derive conclusions on populations $\mathcal{P}(\mathcal{X})$.

Notation:

- $\mathcal{X}$ sample space
- $(\mathcal{X}, \mathcal{P})$ a probability space (nonparametric model)
- $(\mathcal{X}, P(X, \theta) \;\; \theta \in \Theta)$ a probability space (parametric model)
- X^n a random sample of size n
- X random variable prior to observation
- x_i an observed outcome of the random variable

Summary Measures for Sampled Data

Measures to describe or characterize a population

- Measure of central tendency, location
- Measure of dispersion

One such a quantity is

- a **parameter** if it refers to the population (Greek letters)
- a **statistics** if it is an *estimation* of a population parameter from the sample (Latin letters)

Measures of central tendency

- Arithmetic Average (Sample mean)

$$\bar{X} = \frac{\sum x_i}{n}$$

- *Quantile*: value above or below which lie a fractional part of the data (used in nonparametric statistics)
  - Median

$$\mathcal{M} = x_{(n+1)/2}$$

  - Quartile

$$Q_1 = x_{(n+1)/4} \qquad Q_3 = x_{3(n+1)/4}$$

  - q-quantile

    q of data lies below and $1 - q$ lies above

- Mode

  value of relatively great concentration of data
  (*Unimodal vs Multimodal* distributions)

Measure of dispersion

- Sample range

$$R = x_n - x_1$$

- Sample variance

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{X})^2$$

- Standard deviation

$$sd = \sqrt{s}$$

- Inter-quartile range

$$IQR = Q_3 - Q_1$$

Graphical representation of data

Data from a random sample $X^n \sim \mathcal{P}(X)$

- Bar Plots/Histograms (frequency of observations)

```
> hist(r$q.greedy)
> barplot(table(r$q.greedy))
```

- Smoothed density distribution

```
> hist(r$q.greedy,freq=FALSE)
> lines(density(r$q.greedy,bw=1))
```

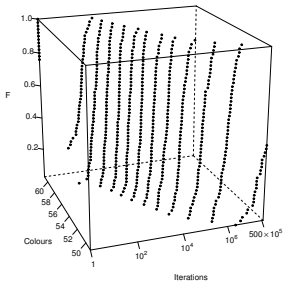
- Empirical Cumulative Distribution Function

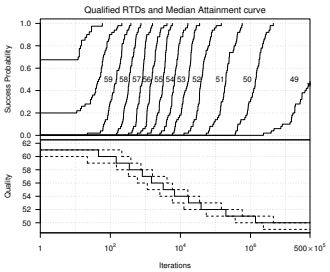
```
> plot.ecdf(r$q.greedy,verticals=T,do.p=F)
> library(Hmisc)
> region <- factor(sample(c("Europe", "USA", "Australia"), 100, TRUE))
> year <- factor(sample(2001:2002, 1000, TRUE))
> ecdf(~ch | region * year, groups = sex)
```

- Boxplots

```
> boxplot(r[,c(1,3,5)])
```

Empirical Qualified Run Time Distributions





Correlation Analysis

Considers data sets that consists of more than one random variable. They are called *multivariate* (or *bivariate* if the variables are two).

- Scatter plot visualization
- Pearson Correlation coefficient

$$r = \frac{1}{n-1}\sum_i x'_i y'_i$$

with

$$x'_i = \frac{x_i - \bar{x}}{s_x} \qquad y'_i = \frac{y_i - \bar{y}}{s_y}$$

Simple Linear Regression

Considers only two variables: *dependent variable* and an *independent variable*

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Uses the Least squares criterion:

$$e_i = y_i - \alpha - \beta x_i$$

$$\min \sum e^2$$

Exploratory Data Analysis

Definition: Exploratory Data Analysis is the process of looking at the data in many different ways in order to get an initial understanding of the phenomenon under study.

Relies on the previously introduced computer graphics and summary measures to suggest interesting questions.

It is *not* meant to

- establish conclusive evidence to answer specific questions, nor to
- generalize beyond the set of data being analyzed

These are the tasks of *Inferential Statistics*.

Outline

1. Descriptive Statistics
   - Computer Graphics for Sampled Data
   - Correlation and Linear Regression
   - Characterization

2. Inferential Statistics
   - Two Sample Tests
   - Analysis of Variance

Motivations

- The concept of optimality can be relatively vague to the management.

  Management is more likely to be convinced if said that the heuristic method outperform significantly the current practice or the currently used method.
- With Metaheuristics results are more volatile as they depends on several detailed choices.

The approach is Methodological:

- adopt tools for taking correct decisions in the development of LSM.

Where do we need inference?

- in the prediction of algorithm results
- in the comparison between algorithms
- in the analysis of the impact of algorithmic factors
- in the goodness of fit of distributions

Inferential Statistics

- We work with samples (instances, solution quality)
- But we want sound conclusions: generalization over a given population (all possible instances)
- Thus we need Statistical Inference

Random Sample X^n
Statistical Estimator $\widehat{\theta}$
$\xrightarrow{\text{Inference}}$
Population $P(x, \theta)$
Parameter θ

Note: Descriptive statistics are nice but one should never infer from a median, average or percentile. Often in the literature: *"the proposed algorithm is better than algorithm X because it gives better average results on some instances (out of a benchmark of 20)"*

Parameter Estimation

Estimator $\hat{\theta}(X_1, \ldots, X_n)$ makes a guess on the parameter (Es. $\bar{X}$)

Estimate is the actual value $\hat{\theta}(x_1, \ldots, x_n)$

Properties of an estimator:

- unbiased: $E[\hat{\theta}] = \theta$ (*e.g.*, $E[\bar{X}] = \mu$)
- consistent
- efficient (uncertainty must decrease with size, *e.g.*, $\text{Var}[\bar{X}] = \sigma^2/n$)
- sufficient

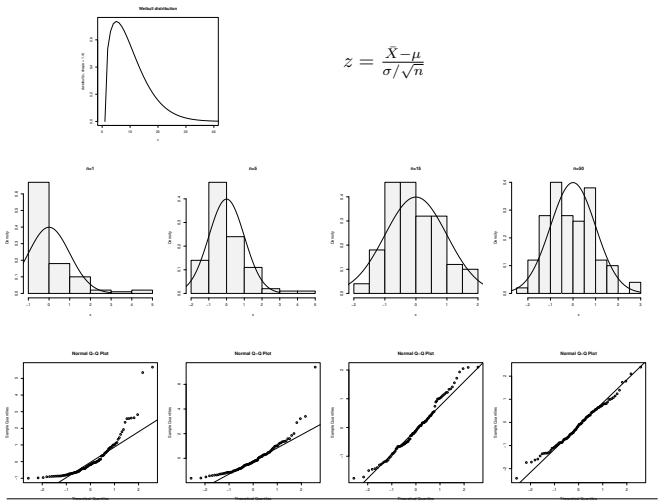
Note: The *best* result $b_N = \min_i c_i$ is not a good estimator. It is biased and not efficient.

Theorem: Central Limit Theorem

If X^n is a random sample from an **arbitrary** distribution with mean μ and variance σ then the average $\bar{X}^n$ is asymptotically normally distributed, *i.e.*,

$$\bar{X}^n \approx N(\mu, \frac{\sigma^2}{n}) \qquad \text{or} \qquad z = \frac{\bar{X}^n - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

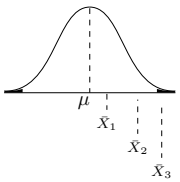
- Consequences:
  - allows inference from a sample
  - allows to model errors in measurements: $X = \mu + \epsilon$
- Issues:
  - n should be *enough* large
  - μ and σ must be known



$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

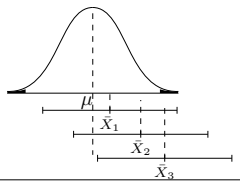
Inference: Hypothesis Testing and Confidence Intervals

A test of hypothesis determines how likely an sampled estimate $\hat{\theta}$ is to occur under some assumptions on the parameter θ of the population.



$$Pr\Big\{\mu - z_1\frac{\delta}{\sqrt{n}} \le \bar{X} \le \mu + z_2\frac{\delta}{\sqrt{n}}\Big\} = 1-\alpha$$

A confidence interval contains all those values that a parameter θ is likely to assume with probability $1-\alpha$: $Pr(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1-\alpha$



$$Pr\Big\{\bar{X} - z_1\frac{\delta}{\sqrt{n}} \le \mu \le \bar{X} + z_2\frac{\delta}{\sqrt{n}}\Big\} = 1-\alpha$$

And if the variance is unknown...

then we substitute σ with its estimator $\hat{\sigma} = S$

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

but then

$$z = \frac{X-\mu}{S\sqrt{n}} \approx t_{n-1}$$

i.e., z approximates a t-student distribution.

Terminology

Statistical Hypotheses:

- H_0: null hypothesis
- H_1 alternative hypothesis (one-sided, two-sided)

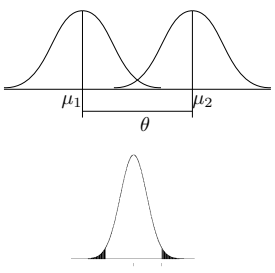
Within the testing procedure two types of errors are possible:

- error of Type I when the null hypothesis is rejected although it is true the level α specifies a priori the assumed likelihood of this error α is called level of significance
- error of Type II occurs when a false null hypothesis is not rejected it is denoted by β

The power of the test is the likelihood of rejecting a false null hypothesis, $1-\beta$.

The power of a test depends on the test statistics, the sample size, the alternative hypothesis.

Hypothesis Testing Procedure



1. Specify the parameter θ and the test hypothesis, *e.g.*,
   $$\theta = \mu_1 - \mu_2 \qquad \begin{cases} H_0 : \theta = 0 \\ H_1 : \theta \neq 0 \end{cases}$$
2. Obtain $P(\theta|\theta = 0)$, the null distribution of θ
3. Compare $\hat{\theta}$ with the upper (in case of one-sided tests) α-quantile of $P(\theta|\theta = 0)$ and accept or reject H_0 according to whether $\hat{\theta}$ is smaller or larger than this value.

Two Sample Matched Pairs Case

It corresponds to one run on various instances design

- Student t test `t.test()`
- Wilcoxon test `wilcox.test()`
- Binomial test `binom.test()`

Two Sample Case

It corresponds to a several runs on one single instance design

- Student t test `t.test()`
- Kruskall Wallis `kruskal.test()`

Two issues:

- Need to test more than two algorithms
- Need to test more than one factor

Multisamples – Analysis of Variance

$H_0: \ \mu_1 = \mu_2 = \mu_3 = \ldots$ $\qquad$ H_1 : {at least one differ}

Applying t-test to all pairs the error of Type I is not α but higher:

$$\alpha_{EX} = 1 - (1 - \alpha)^c$$

Eg, for $\alpha = 0.05$ and $c = 3$ $\alpha_{EX} = 0.14$!

Single Factor Analysis of Variance

$$X_{it} = \mu_i + \epsilon_{it}$$

$$X_{it} = \mu + \alpha_i + \epsilon_{it}$$

$$\sum_{i=1}^{k}\sum_{t=1}^{r}(X_{it} - \bar{X}_{..})^2 = \sum_{i=1}^{k}\sum_{t=1}^{r}(X_{it} - \bar{X}_{i.})^2 + \sum_{i=1}^{k} r(\bar{X}_{i.} - \bar{X}_{..})^2$$

(decomposition in *within-group* and *between-group* sum of squares)

$$MST = \frac{\sum_{i=1}^{k} r\left(\bar{X}_{i.} - \bar{X}_{..}\right)^2}{k-1} \qquad MSE = \frac{\sum_{i=1}^{k}\sum_{t=1}^{r}\left(X_{it} - \bar{X}_{i.}\right)^2}{N-k}$$

(MST *mean square per treatment* and MSE *mean square per error*)

$$F = \frac{MST}{MSE}$$

$F \sim F_{k-1,r-k}$, *i.e.*, the F-ratio approximates a the Fisher distribution F_{df_1,df_2} with df_1 and df_2 degrees of freedom.

Single Factor analysis of Variance (contd.)

- Parametric analysis: ANOVA through F-ratio and Fisher test
  - independent
  - normally distributed
  - homoschedastic
- Nonparametric analysis: Kruskall Wallis
  Assumptions
  - independent
  - homoschedastic

Two Factors Analysis of Variance

Model:

$$X_{hi} = \mu + \alpha_i + \theta_h + \epsilon_{hi}$$

Mean square:

$$MST = \frac{b\sum_{i=1}^{k}(\bar{X}_{.i} - \bar{X}_{..})^2}{k-1}; \quad MSE = \frac{\sum_{h=1}^{b}\sum_{i=1}^{k}\left(X_{hi} - \bar{X}_{h.} - \bar{X}_{.i} + \bar{X}_{..}\right)^2}{bk - b - k + 1}$$

Statistical Tests

- ANOVA through F-ratio and Fisher test
- Friedman test

Two Factors Repeated Measures Analysis of Variance

Model:

$$X_{hit} = \mu + \alpha_i + \theta_h + \epsilon_{hit}$$

or alternatively:

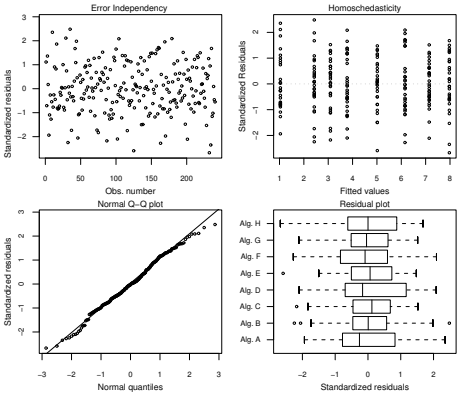
$$X_{hit} = \mu + \alpha_i + \theta_h + \alpha\theta_{hi} + \epsilon_{hit}$$

Statistical Tests

- ANOVA through F-ratio and Fisher test
- Friedman test

If the interactions between factors are of interest interaction plots are useful to visualize them

ANOVA Assumptions

- Parametric analysis: ANOVA through F-ratio and Fisher test
  - independent
  - normally distributed
  - homoschedastic
- Nonparametric analysis: through Rank based tests
  - independent
  - homoschedastic

Check Assumptions