LOCAL SEARCH METHODS
APPLICATIONS AND ENGINEERING

Lecture 11

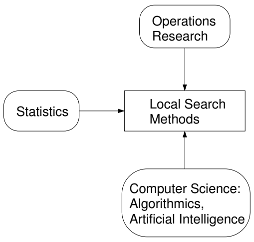
Empirical Methods
Introduction to Statistics

Marco Chiarandini

Outline

Outline

The use of Local Search Methods as an interdisciplinary activity



Statistics: Analysis and interpretation of data with a view toward objective evaluation of the replicability of the conclusions based on the data.
Field of mathematics that studies the probability of events on the basis of inference from empirical data.

Descriptive statistics: Resume and visualize data (Exploratory data analysis)

Inferential statistics: make inference or prediction about the populations from which samples are drawn.

Population: total of subjects that share something in common
Sample: set of subjects drawn from populations

Data:

- quantitative (numerical) discrete or continuous (presence of an order)
- qualitative or categorical

Why Statistics in LS Methods?

Statistics deals with *random (or stochastic) variables*.

A variable is called random if, prior to observation, its outcome cannot be predicted with certainty.

The uncertainty is described by a *Probability Distribution*.

In the analysis of LS algorithms:

- A class of instances $\mathcal{I}$ is made by *a priori* indistinguishable instances I. They constitute a population with probability distribution $\mathcal{P}_I$.
- Given an instance, the solution returned by an algorithm is a stochastic quantity, *i.e.*, with random probability distribution $\mathcal{P}_c$

Hence, the performance is determined by two stochastic variables.
In fact, things are even more complex: time is also a stochastic variable.

Reasons for statistical studies on LS algorithms

- Tuning:
  which is the best set of algorithm parameters?
- Comparisons for decision making:
  is algorithm A better than B?
- Characterization and prediction:
  how does the alg. perform in average and how does it scale?
- Analysis of impact of algorithm components:
  is the introduction of a new component really effective? does it interacts with other components?
- Learning?

Performance Measures

- Decision problems: computational time (or # iterations)
- Optimization problems:
  - computational time
  - solution quality

  Simplification: study only sol. quality but use *Fairness principle*, *i.e.*, allow all to use the same computational resources

Which measure for solution quality?

- Distance or error from optimal value
  - optimal value computed exactly or known by instance construction
  - surrogate value such bounds or best known values
  - instances imply different scales hence needed a normalization

$$z(c) = \frac{c - \bar{c}}{s} \quad \text{standard score} \tag{1}$$

$$e_1(c) = \frac{|c - c_{opt}|}{c_{opt}} \quad \text{relative error} \tag{2}$$

$$e_2(c) = \frac{c - c_{opt}}{c' - c_{opt}} \tag{3}$$

- Rank (no need for normalization but loose of information)

Outline

Random Variables and Probability

Discrete variables

Probability distribution:

$$p_i = P[x = v_i]$$

Cumulative Distribution Function (CDF)

$$F(v) = P[x \leq v] = \sum_i p_i$$

Mean

$$\mu = E[X] = \sum x_i p_i$$

Variance

$$\sigma^2 = E[(X - \mu)^2] = \sum (x_i - \mu)^2 p_i$$

Continuous variables

Probability density function (pdf):

$$f(v) = \frac{dF(v)}{v}$$

Cumulative Distribution Function (CDF):

$$F(v) = \int_{-\infty}^{v} f(v)dv$$

Mean

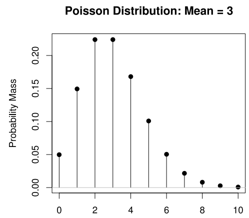
$$\mu = E[X] = \int x f(x) dx$$

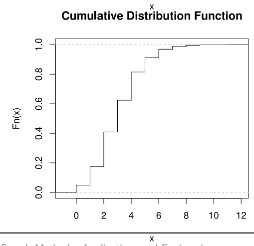
Variance

$$\sigma^2 = E[(X-\mu)^2] = \int (x-\mu)^2 f(x) dx$$

Discrete variables

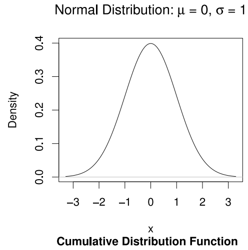
$$p(x) = \frac{e^{\mu}\mu^{x}}{x!} \text{(binomial)}$$

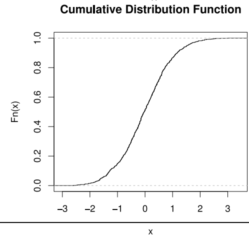




Continuous variables

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \text{(normal)}$$

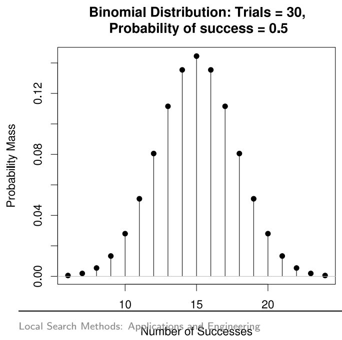




Probability Distributions

Binomial distribution

$$P[x = v] = \binom{n}{v} p^v (1-p)^{n-v}$$

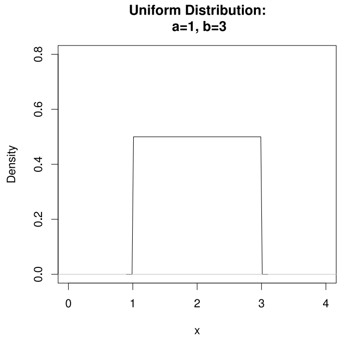


p probability of successes
x number of successes
The binomial distribution indicates the probability for each set of outcomes, *i.e.*, $v = \{1, \ldots, n\}$ successes.

One parameter: p

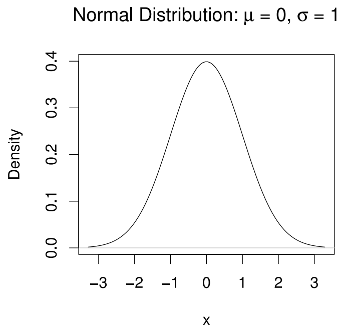
Uniform distribution (continuous)

$$f(x) = \frac{1}{b - a}$$



Normal distribution (continuous)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



Theoretical importance

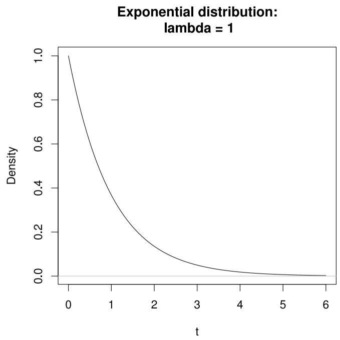
Defined by two parameters: $N(\mu, \sigma)$.

$N(0, 1)$ is the standardized version.

In $N(0, 1)$ 68.27% of data fall within $\mu \pm \sigma$

Exponential distribution (continuous)

$$f(t) = \lambda e^{-\lambda t}$$

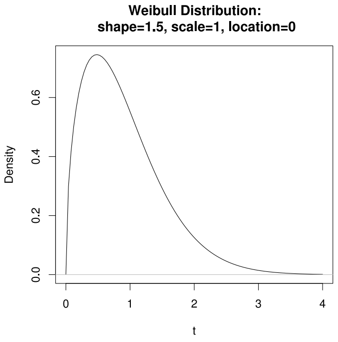


It has the memory-less property, *i.e.*, the probability of a new event to happen within a fixed time does not depend on the time passed so far.

Defined by one parameter: $E[X] = \frac{1}{\lambda}$.

Weibull distribution (continuous)

$$f(x) = \frac{\beta}{\eta}\Big(\frac{t-\gamma}{\eta}\Big)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^{\beta}}$$



Used in life data and reliability analysis

Defined by three parameters:
β (shape), η (scale), γ (location)

Others (theoretically relevant)

- $\chi^2(n)$: chi-squared distribution with n degrees of freedom:
  distribution of $\sum_i X_n^2$ where $X_1, \ldots, X_n$ are independently, standard normally distributed variables
- $t(r)$: Student t-distribution with r degrees of freedom:
  distribution of $X_1/\sqrt{X_2/r}$ with $X_1 \sim N(0,1)$ and $X_2 \sim \chi^2(r)$ independently distributed variables
- $F(r_1, r_2)$: Fisher distribution with r_1 and r_2 degrees of freedom:
  distribution of $(X_1/r_1)/(X_2/r_2)$ with $X_1 \sim \chi^2$ and $X_2 \sim \chi^2$ independently distributed variables

Outline

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

$$s = \sqrt{s}$$

- 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))
```

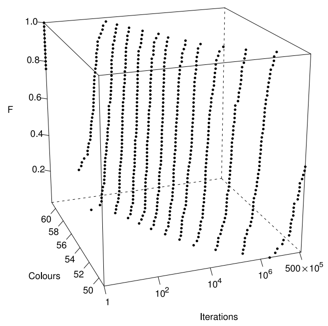
- Empirical Cumulative Distribution Function

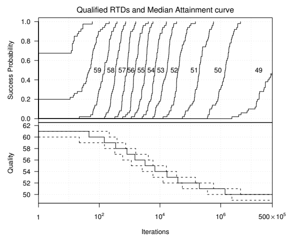
```
> plot.ecdf(r$q.greedy,verticals=T,do.p=F)
> library(Hmisc)
> region <- factor(sample(c("Europe", "USA", "Australia"), 100, TRUE)
> year <- factor(sample(2001:2002, 1000, TRUE))
> ecdf(~ch | region * year, groups = sex)
```

- 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*.