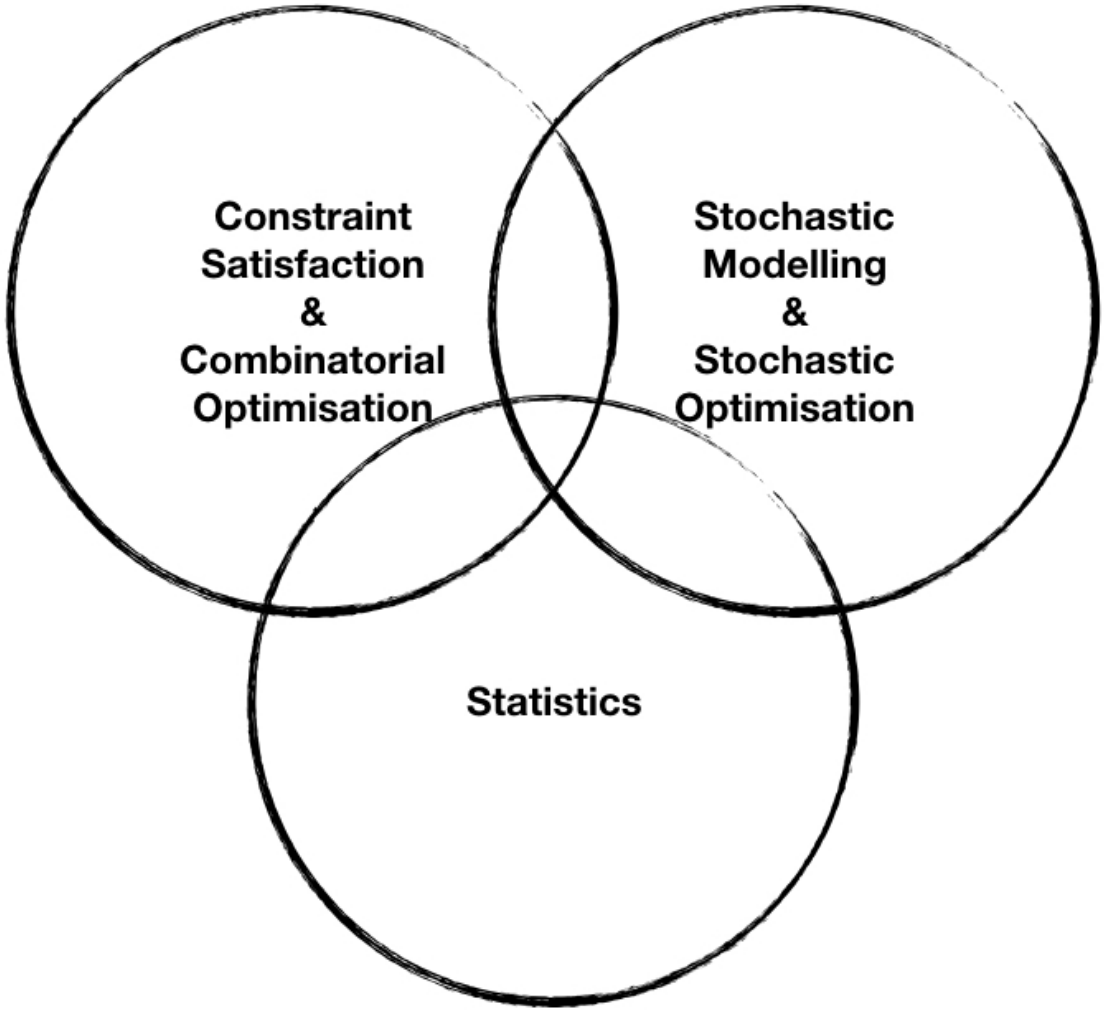
Declarative Statistics

Roberto Rossi,[1] Özgür Akgün,[2] Steven D. Prestwich,[3]
S. Armagan Tarim[4]

[1]The University of Edinburgh Business School, The University of Edinburgh, UK
[2]Department of Computer Science, University of Saint Andrews, UK
[3]Insight Centre for Data Analytics, University College Cork, Ireland
[4]Business School, University College Cork, Ireland

LERIA
Laboratoire d'Etude et de Recherche en Informatique d'Angers
28th June 2019

Context


Constraint
Satisfaction
&
Combinatorial
Optimisation
Stochastic
Modelling
&
Stochastic
Optimisation
Statistics

ConstraintProgramming/Combinatorial Optimisation

Sudoku

5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

Sudoku

Modeling strategies

How do we model (and solve) this?

Sudoku

Constraint Programming model in OPL

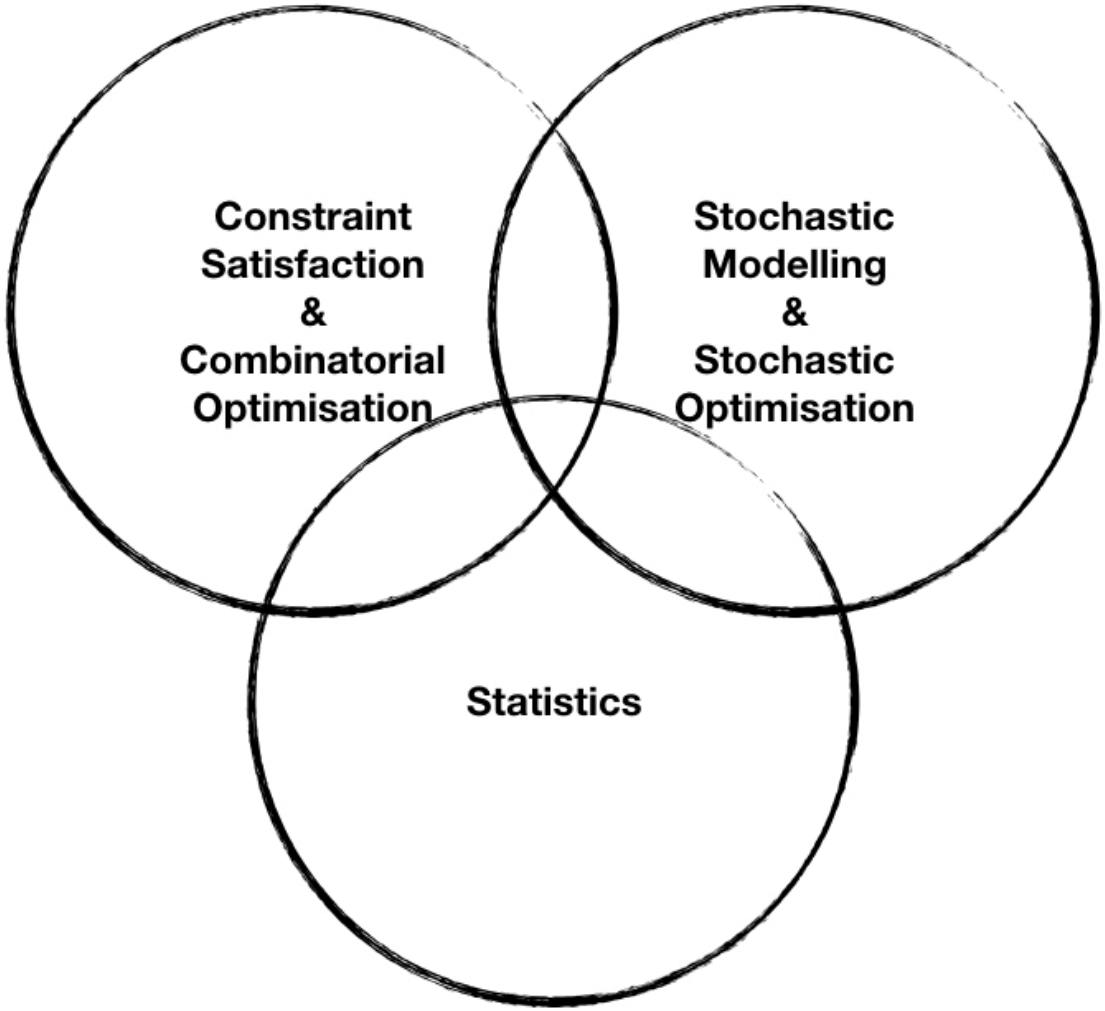
```
using CP;
int grid[1..9][1..9] = ...; // initial grid

dvar int x[1..9][1..9] in 1..9; // unfixed variables

subject to {
  forall (i in 1..9, j in 1..9)
  if (grid[i,j] != 0)
    grid[i,j] == x[i,j]; // fix variables

  forall (idx in 1..9)
    allDifferent(all (i in 1..9) x[idx,i]) && // horizontal
    allDifferent(all (i in 1..9) x[i,idx]) && // vertical
    allDifferent(all (i in 1..3, j in 1..3) x[((idx-1) div 3)*3 +
i][((idx-1) % 3)*3 + j]); // 3x3
};
```

Context


Constraint
Satisfaction
&
Combinatorial
Optimisation
Stochastic
Modelling
&
Stochastic
Optimisation
Statistics

Given a random demand d_t with probability density function f_t in each period, ordering cost

$$c(Q) = \begin{cases} K + vQ & Q > 0 \\ 0 & \text{otherwise} \end{cases},$$

holding cost h and penalty cost p, determine the optimal replenishment strategy over a T period planning horizon.

Stochastic lot sizing

Modeling strategies

How do we model (and solve) this?

Stochastic lot sizing

Stochastic (Dynamic) Programming

$$L(y) = \min \underbrace{h \int_0^y (y-i) f(i) \mathrm{d}i}_{\text{holding cost}} + \underbrace{p \int_y^\infty (i-y) f(i) \mathrm{d}i}_{\text{penalty cost}}$$

$$C_n(x) = \min_Q \left\{ c(Q) + L(Q+x) + \int_0^\infty C_{n-1}(x+Q-i) f(i) \mathrm{d}i \right\}$$

$$G_n(y) = L(y) + \int_0^\infty C_{n-1}(y-i) f(i) \mathrm{d}i$$

where $C_1(I) = \min_Q \{c(Q) + L(Q+I)\}$.

Java: `http://gwr3n.github.io/jsdp/`
Python: `https://pypi.org/project/inventoryanalytics/`

Stochastic lot sizing

Stochastic (Dynamic) Programming

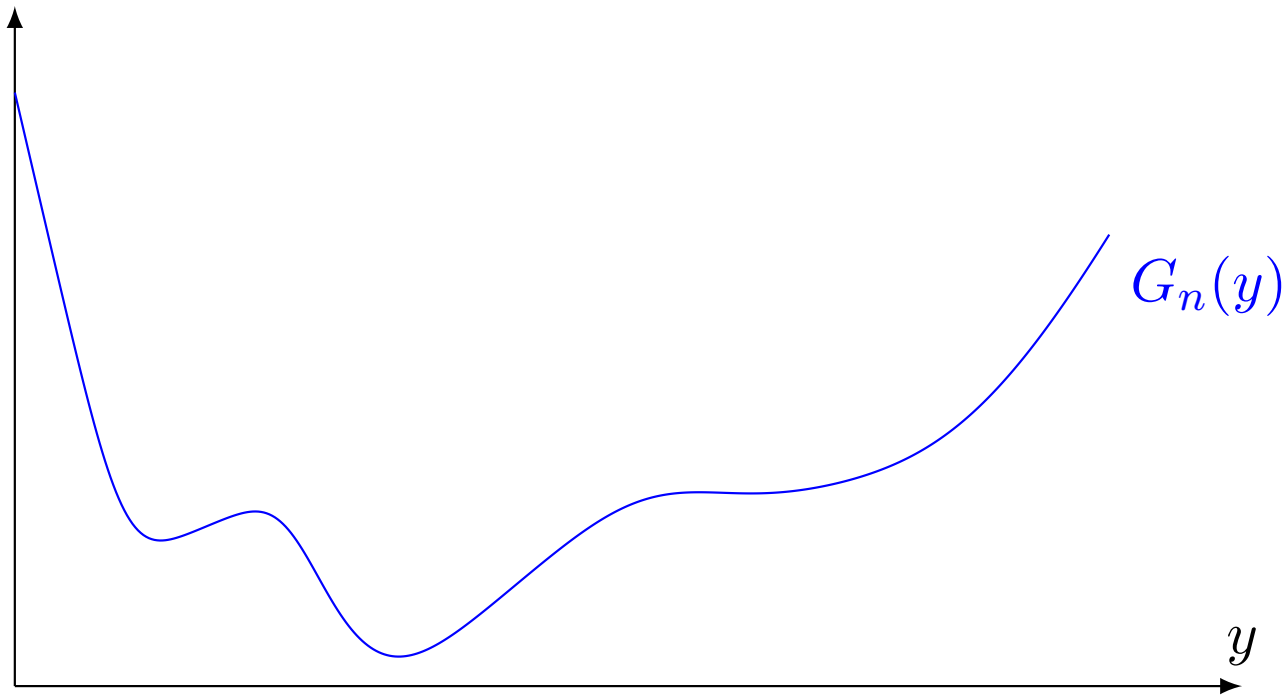
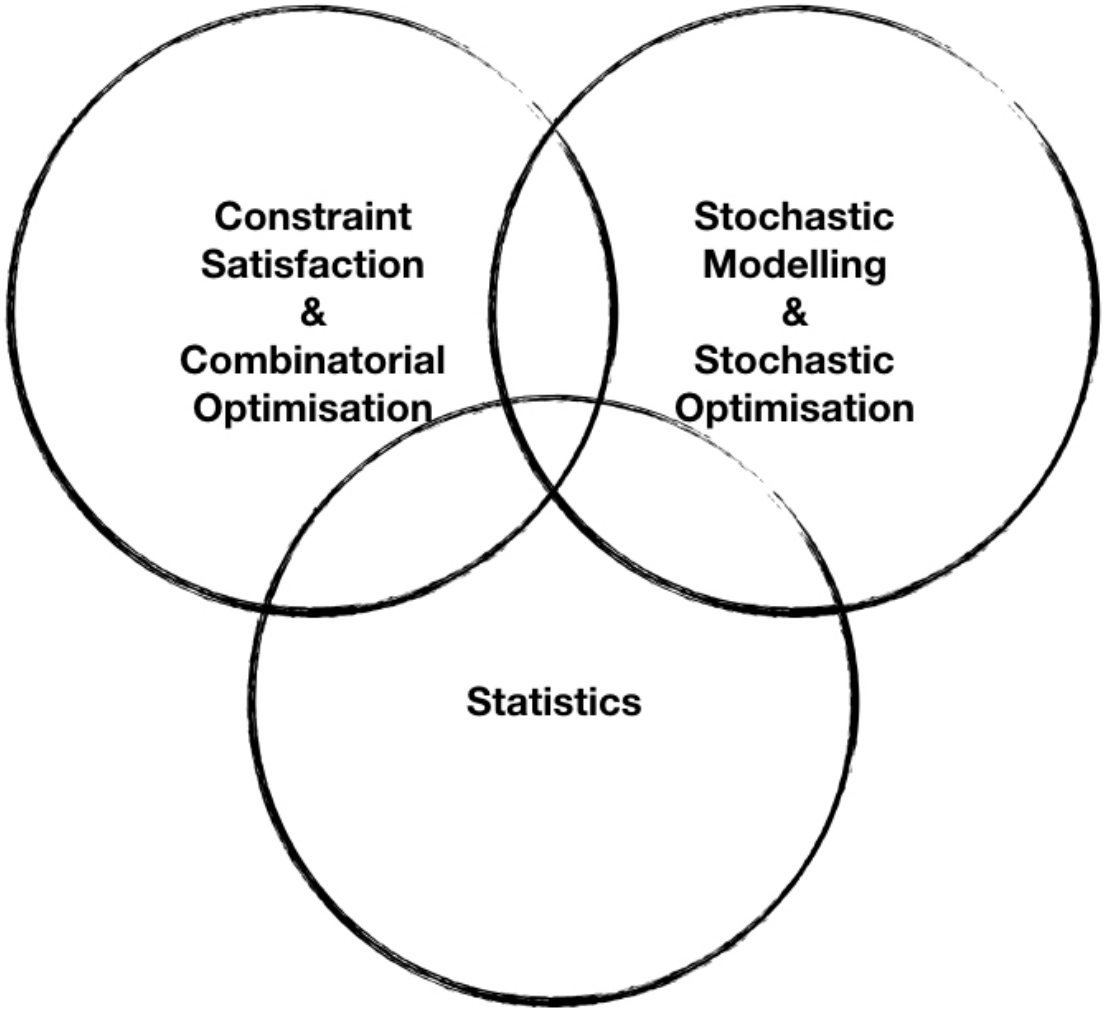


Figure: Plot of $G_n(y)$

Context


Constraint
Satisfaction
&
Combinatorial
Optimisation
Stochastic
Modelling
&
Stochastic
Optimisation
Statistics

Statistics

Comparing sample mean across groups

Group 1	Group 2	Group 3
3.57329	9.83132	9.80335
6.5655	9.7379	8.79726
-2.06033	6.6339	13.6045
0.469477	8.20049	9.4932
3.05632	7.19737	8.50685
5.54063	9.19586	9.22433

Table: Groups considered in our example.

Comparing sample mean across groups

Modeling strategies

How do we model (and solve) this?

Comparing sample mean across groups

ANOVA

Mathematica: `ANOVA` command.

		DF	SumOfSq	MeanSq	F-Ratio	p-value
ANOVA	Model	2	166.387	83.1935	16.0089	0.000189984
	Error	15	77.9505	5.1967		
	Total	17	244.337			
Means	Overall	7.07618				
	Group 1	2.85748				
	Group 2	8.46614				
	Group 3	9.90492				

Table: One-way analysis of variance table; DF: degrees of freedom; SumOfSq: sum of squared differences; MeanSq: mean squared differences.

Comparing sample mean across groups

Problem

Lack of a "declarative" modelling framework!

Comparing sample mean across groups

What about "Statistical Programming"?

Constraints:

(1) $\mu_i = \mu_j$ for all i, j where $i < j$

(2) $t^2(\langle O \rangle; \mu; \alpha)$

Parameters:

m	number of groups
n	number of random variates within a group
$O_{i,j}$	random variate j in group i
α	significance level

Decision variables:

μ_i	mean of group i

Figure: Statistical Programming model for testing equality of means (infeasible).

Comparing sample mean across groups

Confidence interval analysis

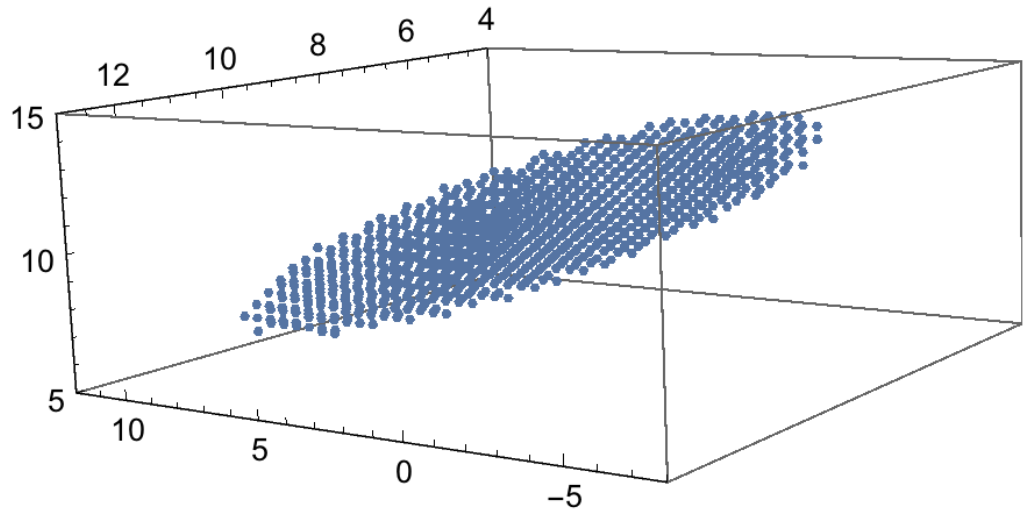


Figure: Confidence region as solution set of a Statistical Program

Comparing sample mean across groups

What about "Statistical Programming"?

Objective:

$\min$ (or $\max$) μ_i

Constraints:

(2) $t^2(\langle O \rangle; \mu; \alpha)$

Parameters:

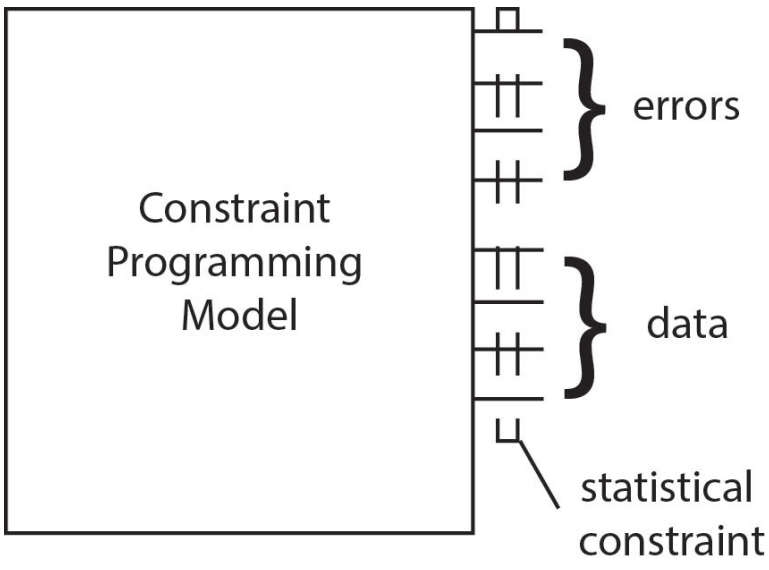
m	number of groups
n	number of random variates within a group
$O_{i,j}$	random variate j in group i
α	significance level

Decision variables:

μ_i	mean of group i

Figure: Statistical Programming model for confidence interval analysis.

Statistical Programming



Constraints:

(1) $\mu_i = \mu_j$ for all i, j where $i < j$

(2) $t^2(\langle O \rangle; \mu; \alpha)$ ←statistical constraint

Parameters:

m	number of groups
n	number of random variates within a group
$O_{i,j}$	random variate j in group i
α	significance level

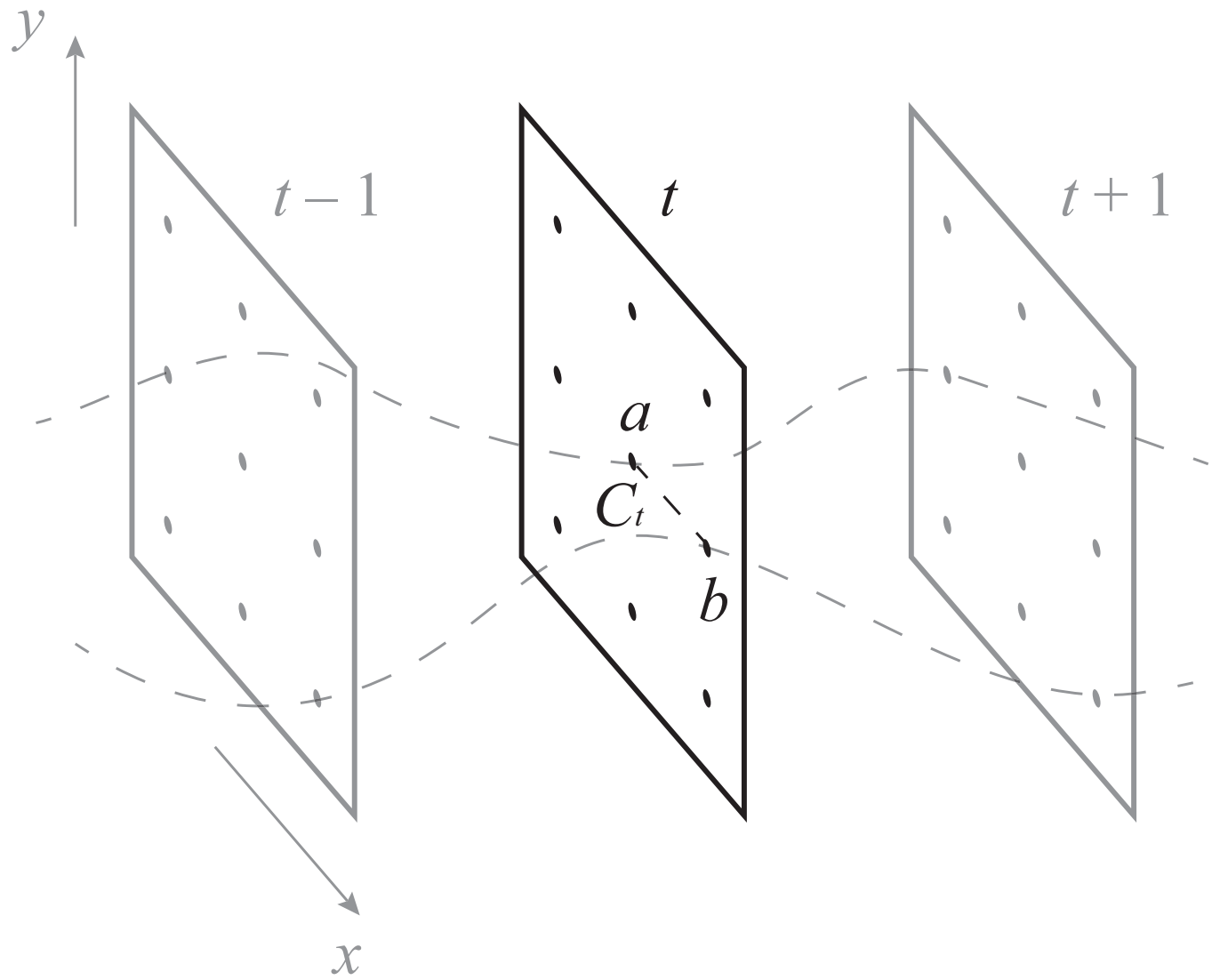
Decision variables:

μ_i	mean of group i

What is a **statistical model**?

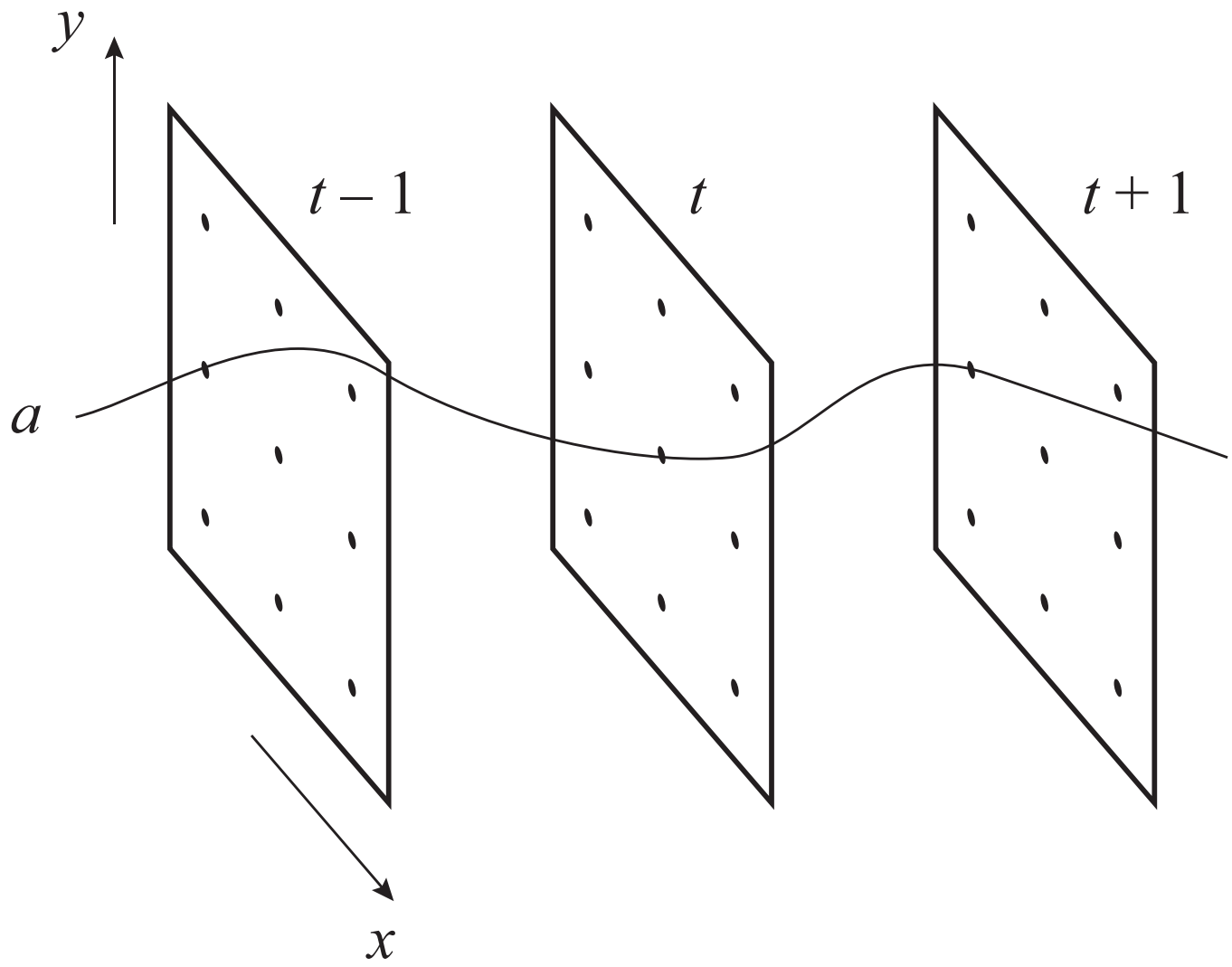
Statistics

Cross sectional study



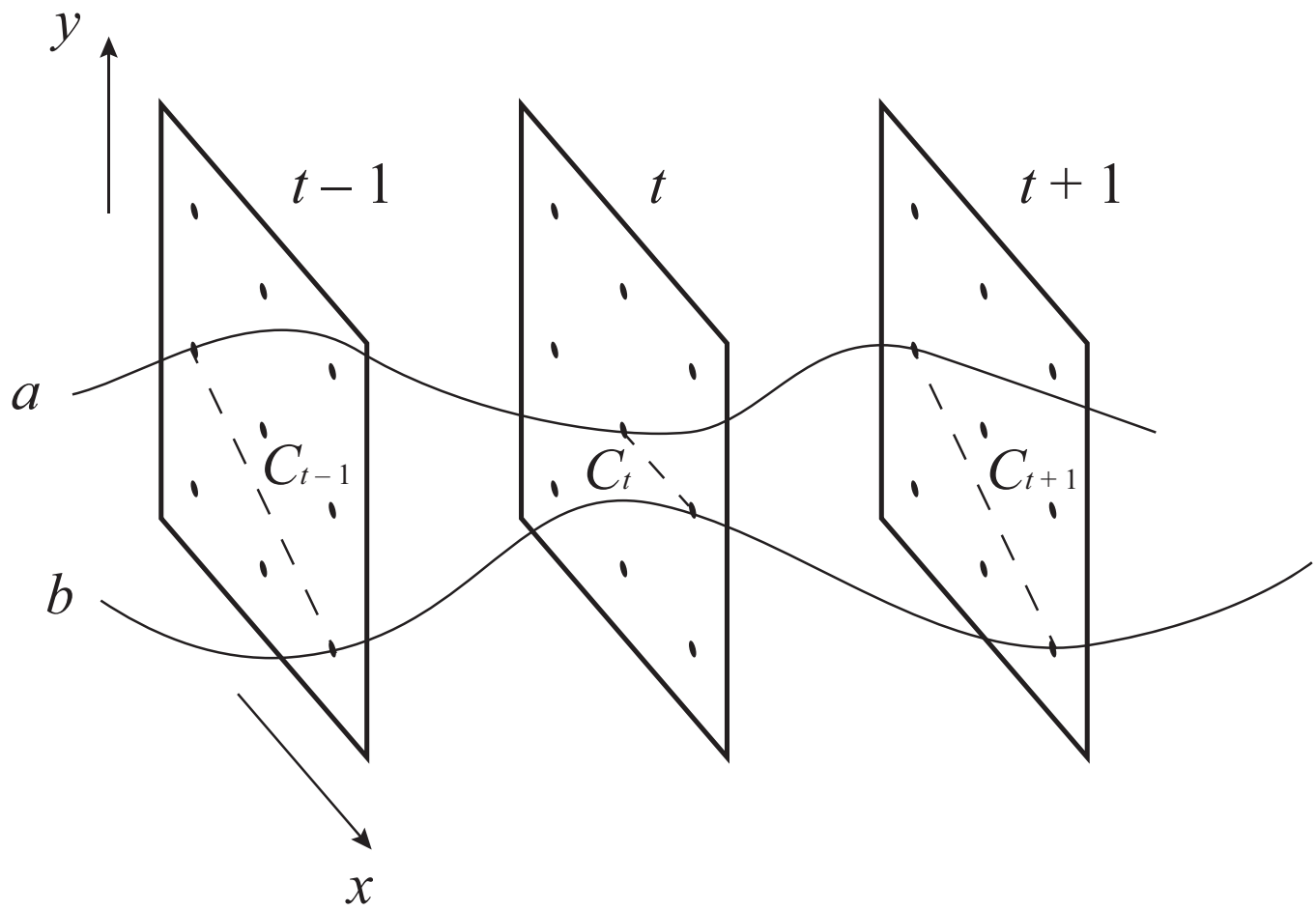
Statistics

Time series analysis



Statistics

Panel data



Statistics

Panel data

pandas (software)

From Wikipedia, the free encyclopedia

This article **relies too much on references to primary sources**. Please improve this by adding secondary or ter

(Learn how and when to remove this template message)

In computer programming, **pandas** is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.[2] The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.[3]

Contents [hide]

1 Library features
2 History
3 See also
4 References
5 Further reading
6 External links

What is a **statistical model**?

P. McCullagh, "What is a statistical model?," The Annals of Statistics, 30(5), pp. 1225—1267, (2002).

Probability theory

Formal background

A **probability space** is a mathematical tool that aims at modelling a **real-world experiment** consisting of outcomes that occur randomly.

It is described by a triple $(\Omega, \mathcal{F}, \mathcal{P})$

- Ω denotes the **sample space** — the set of all possible **outcomes** of the experiment,
- $\mathcal{F}$ denotes the **sigma-algebra** on Ω — i.e. the power set 2^{Ω} of all possible **events** on the sample space
- $\mathcal{P}$ denotes the **probability measure** — i.e. a function $\mathcal{P} : \mathcal{F} \to [0, 1]$ returning the probability of each possible event.

Probability theory

Random variable

A **random variable** ω is an $\mathcal{F}$-measurable function $\omega : \Omega \to \mathbb{R}$ defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ mapping its sample space to the set of all real numbers.

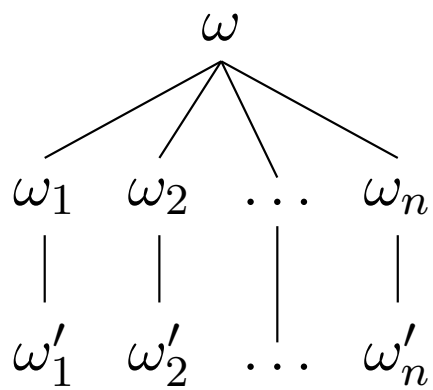
Given ω, we can ask questions such as "what is the probability that ω is less or equal to element $s \in \mathbb{R}$."

This is the probability of event $\{o : \omega(o) \leq s\} \in \mathcal{F}$, which is often written as $F_\omega(s) = \Pr(\omega \leq s)$, where $F_\omega(s)$ is the **cumulative distribution function** (CDF) of ω.

Probability theory

Multivariate random variable

A **multivariate random variable** is a random vector $(\omega_1, \ldots, \omega_n)^T$, where ^{T} denotes the "transpose" operator.



The **outcome** of the experiment is the vector of **random variates** $(\omega'_1, \ldots, \omega'_n)^T$, which are **scalars**.

Criticism: This experiment is lost in classical "procedural" statistics (e.g. `ANOVA` command) because commands are black-box!

Statistical Model

Definition

Consider a **multivariate random variable** defined on probability space $(\Omega, \mathcal{F}, \mathcal{P})$

Let $\mathcal{D}$ be **a set of possible CDFs** on the sample space Ω.

We adopt the following definition of a statistical model

Definition

A statistical model is a pair $\langle \mathcal{D}, \Omega \rangle$.

Nonparametric Statistical Model

Definition

Let $\mathbb{D}$ denote the set of all possible CDFs on Ω.

Definition

A nonparametric statistical model is a pair $\langle \mathbb{D}, \Omega \rangle$.

Parametric Statistical Model

Definition

Let $\mathbb{D}$ denote the set of all possible CDFs on Ω.

Consider a finite-dimensional parameter set Θ together with a function $g : \Theta \to \mathbb{D}$, which assigns to each parameter point $\theta \in \Theta$ a CDF F_θ on Ω.

Definition

A parametric statistical model is a triple $\langle \Theta, g, \Omega \rangle$.

Statistical Inference

Definition

Consider now the **outcome** $o \in \Omega$ of an experiment.

Statistics operates under the assumption that there is a **distinct element** $d \in \mathcal{D}$ that generates the observed data o.

The aim of statistical inference is then to determine which element(s) are **likely to be the one** generating the data.

Hypothesis Testing

Modus operandi 1 of 2

In hypothesis testing the statistician selects a **significance level** α and formulates a **null hypothesis** (H_0), e.g.

"element $d \in \mathcal{D}$ has generated the observed data,"

and an **alternative hypothesis**, e.g.

"another element in $\mathcal{D}/d$ has generated the observed data."

Hypothesis Testing

Modus operandi 2 of 2

She selects a **statistical test** and derive the **distribution of the associated test statistic** under the null hypothesis.

By using this distribution, she determines the probability of obtaining **a test statistic at least as extreme** as the one associated with outcome o.

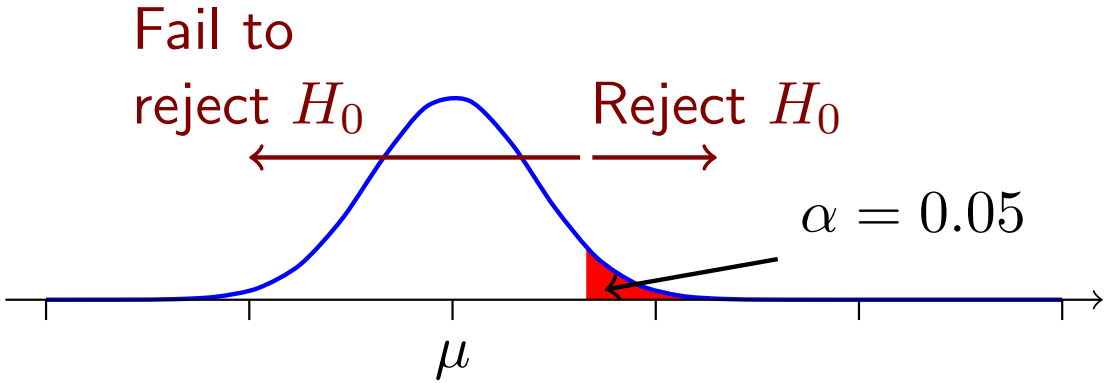
If this probability is less than α: "reject the null hypothesis." Observed result is highly unlikely under the null hypothesis.

If this probability is greater or equal to α: "fails to reject the null hypothesis." Evidence insufficient to support a conclusion against the null hypothesis.

Parametric Hypothesis Testing

Student's t-test (one-tailed)

Student's t-test statistics, $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$, follows a T distribution with $n-1$ degrees of freedom.

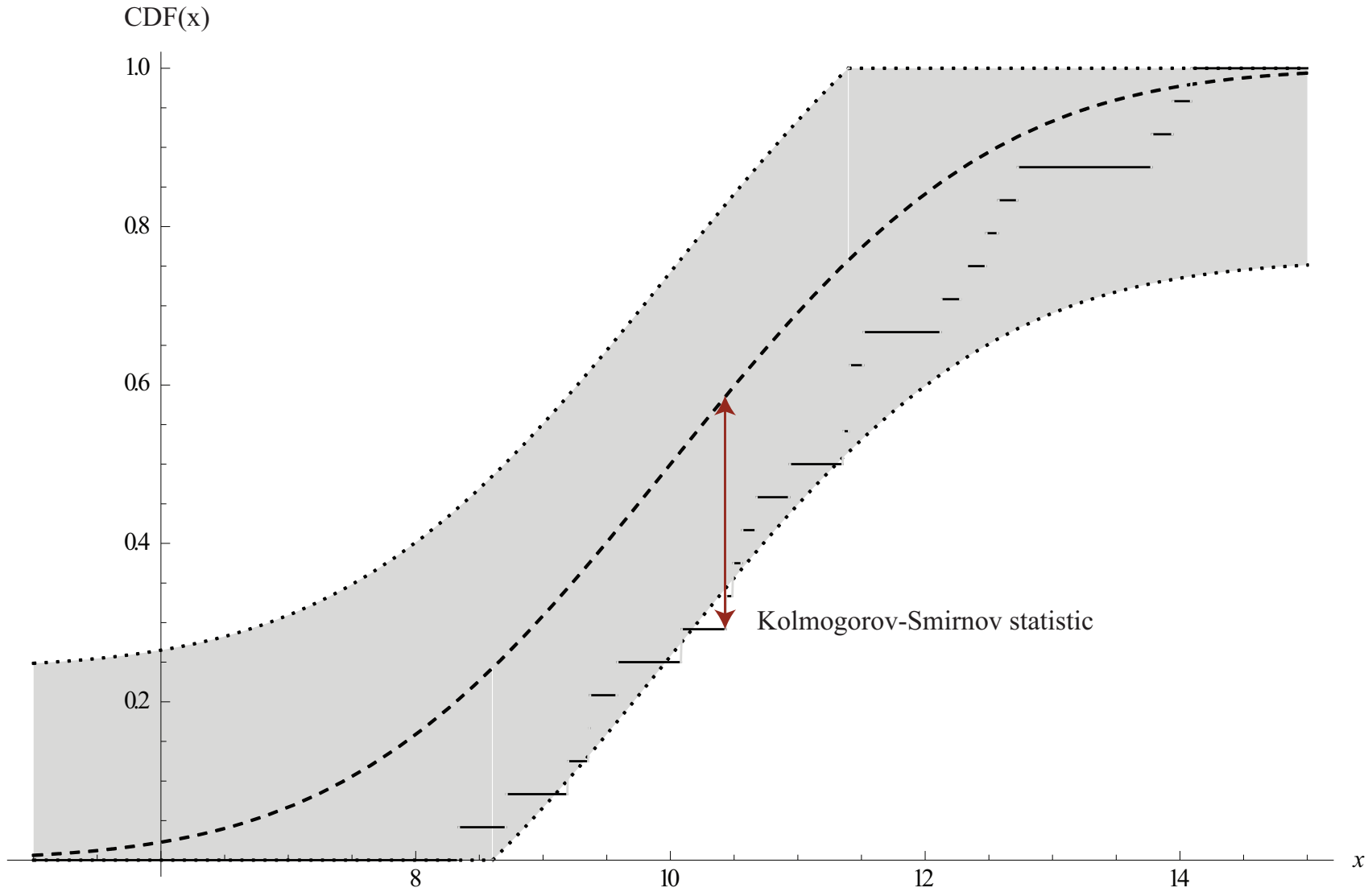


The one-sample t-test compares the mean of a sample to value μ.

The two-sample t-test compares means μ_1 and μ_2 of two samples.

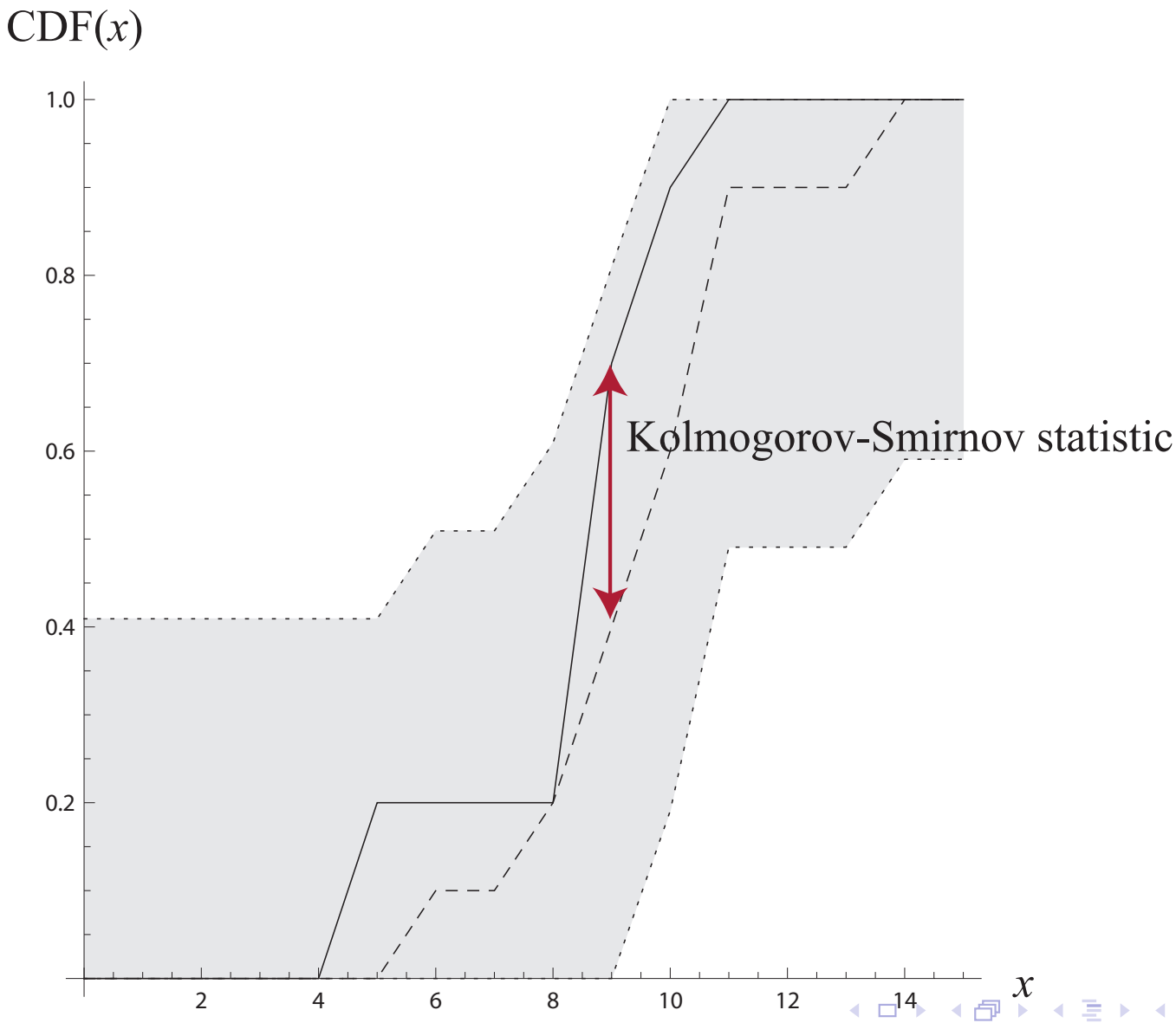
Parametric Hypothesis Testing

One-sample Kolmogorov-Smirnov test (two-tailed)



Nonparametric Hypothesis Testing

Two-sample Kolmogorov-Smirnov test (two-tailed)



Statistical Constraints

Definition

A **statistical constraint** is a constraint that embeds **a parametric or a non-parametric statistical model** and a **statistical test** with **significance level** α that is used to determine which assignments satisfy the constraint.

ECAI 2014
T. Schaub et al. (Eds.)
© 2014 The Authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-419-0-777

Statistical Constraints

Roberto Rossi[1] and **Steven Prestwich**[2] and **S. Armagan Tarim**[3]

Abstract. We introduce statistical constraints, a declarative modelling tool that links statistics and constraint programming. We discuss two statistical constraints and some associated filtering algorithms. Finally, we illustrate applications to standard problems encountered in statistics and to a novel inspection scheduling problem in which the aim is to find inspection plans with desirable statistical properties.

element $s \in \mathbb{R}$." This is the probability of event $\{o : \omega(o) \leq s\} \in \mathcal{F}$, which is often written as $F_\omega(s) = \Pr(\omega \leq s)$, where $F_\omega(s)$ is the cumulative distribution function (CDF) of ω. A multivariate random variable is a random vector $(\omega_1, \ldots, \omega_n)^T$, where ^{T} denotes the "transpose" operator. If $\omega_1, \ldots, \omega_n$ are independent and identically distributed (iid) random variables, the random vector may be used to represent an experiment repeated n times, i.e. a sample, where each replica i generates a random variate ω'_i and the outcome of the ex-

Statistical Constraints

Parametric statistical constraint

A **parametric statistical constraint** c takes the general form

$$c(T, g, O, \alpha)$$

where T and O are sets of **decision variables** and $g : \Theta \to \mathbb{D}$.

Let $T \equiv \{t_1, \ldots, t_{|T|}\}$, then $\Theta = D(t_1) \times \ldots \times D(t_{|T|})$.

Let $O \equiv \{o_1, \ldots, o_{|O|}\}$, then $\Omega = D(o_1) \times \ldots \times D(o_{|O|})$.

An assignment is **consistent** with respect to c if the statistical test fails to reject the associated null hypothesis, e.g. "F_θ generated $o_1, \ldots, o_{|O|}$," at significance level α.

Statistical Constraints

Nonparametric statistical constraint

A **nonparametric statistical constraint** c takes the general form

$$c(O_1, \ldots, O_k, \alpha)$$

where $O_1, \ldots, O_k$ are sets of **decision variables**.

Let $O_i \equiv \{o_1^i, \ldots, o_{|O_i|}^i\}$, then $\Omega = \bigcup_{i=1}^{k} D(o_1^i) \times \ldots \times D(o_{|O_i|}^i)$.

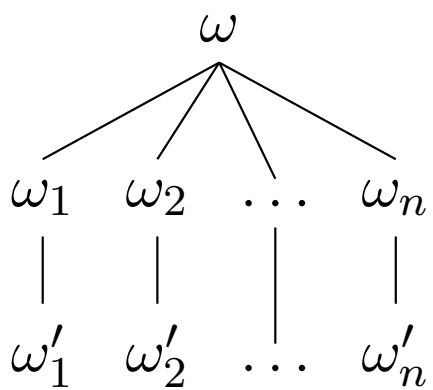
An assignment is **consistent** with respect to c if the statistical test fails to reject the associated null hypothesis, e.g "$\{o_1^1, \ldots, o_{|O_1|}^1\}, \ldots, \{o_1^k, \ldots, o_{|O_k|}^k\}$ are drawn from the same distribution," at significance level α.

Statistical Constraints

Remarks

In statistical constraints, **random variates**, i.e. random variable realisations $(\omega'_1, \ldots, \omega'_n)^T$, are modelled as **decision variables**.

The **sample**, i.e. the set of random variables $(\omega_1, \ldots, \omega_n)^T$ that generated the random variates **is not explicitly modelled**.



Statistical Constraints

Student's t test constraint

$$t\text{-test}_w^{\alpha}(O, m)$$

Statistical Constraints

two-sample Student's t test constraint

$$t\text{-test}_w^{\alpha}(O_1, O_2)$$

Statistical Constraints

Kolmogorov-Smirnov constraint

$$\text{KS-test}_{w}^{\alpha}(O, \text{exponential}(\lambda))$$

Statistical Constraints

Kolmogorov-Smirnov constraint

$$\text{KS-test}_{w}^{\alpha}(O, \text{exponential}(\lambda))$$

Statistical Constraints

two-sample Kolmogorov-Smirnov constraint

$$\text{KS-test}_w^{\alpha}(O_1, O_2)$$

Applications

Classical problems in statistics

Constraints:

(1) $t\text{-test}^{\alpha}_{=}(O, m)$

Decision variables:

$o_1 \in \{8\}, o_2 \in \{14\}, o_3 \in \{6\}, o_4 \in \{12\}, o_5 \in \{12\},$
$o_6 \in \{9\}, o_7 \in \{10\}, o_8 \in \{9\}, o_9 \in \{10\}, o_{10} \in \{5\}$
$O_1 \equiv \{o_1, \ldots, o_{10}\}$
$m \in \{0, \ldots, 20\} \qquad \alpha = 0.05$

Figure: Determining the likely values of the mean of the random variable that generated random variates O_1

After propagating constraint (1), the domain of m reduces to $\{8, 9, 10, 11\}$, so at significance level $\alpha = 0.05$ we reject the null hypothesis that the true mean is outside this range.

Applications

Classical problems in statistics

Constraints:

(1) KS-test$^{\alpha}_{=}(O_1, O_2)$

Decision variables:

$o_1 \in \{9\}, o_2 \in \{10\}, o_3 \in \{9\}, o_4 \in \{6\}, o_5 \in \{11\},$
$o_6 \in \{8\}, o_7 \in \{10\}, o_8 \in \{11\}, o_9 \in \{14\}, o_{10} \in \{11\},$
$o_{11}, o_{12} \in \{5\}, o_{13}, \ldots, o_{20} \in \{9, 10, 11\}$
$O_1 \equiv \{o_1, \ldots, o_{10}\}, O_2 \equiv \{o_{11}, \ldots, o_{20}\} \qquad \alpha = 0.05$

Figure: Devising sets of random variates that are likely to be generated from the same random variable that generated a reference set of random variates O_1

By finding all solutions to the above CSP we verified that there are 365 sets of random variates for which the null hypothesis is rejected at significance level α.

Applications

Classical problems in statistics

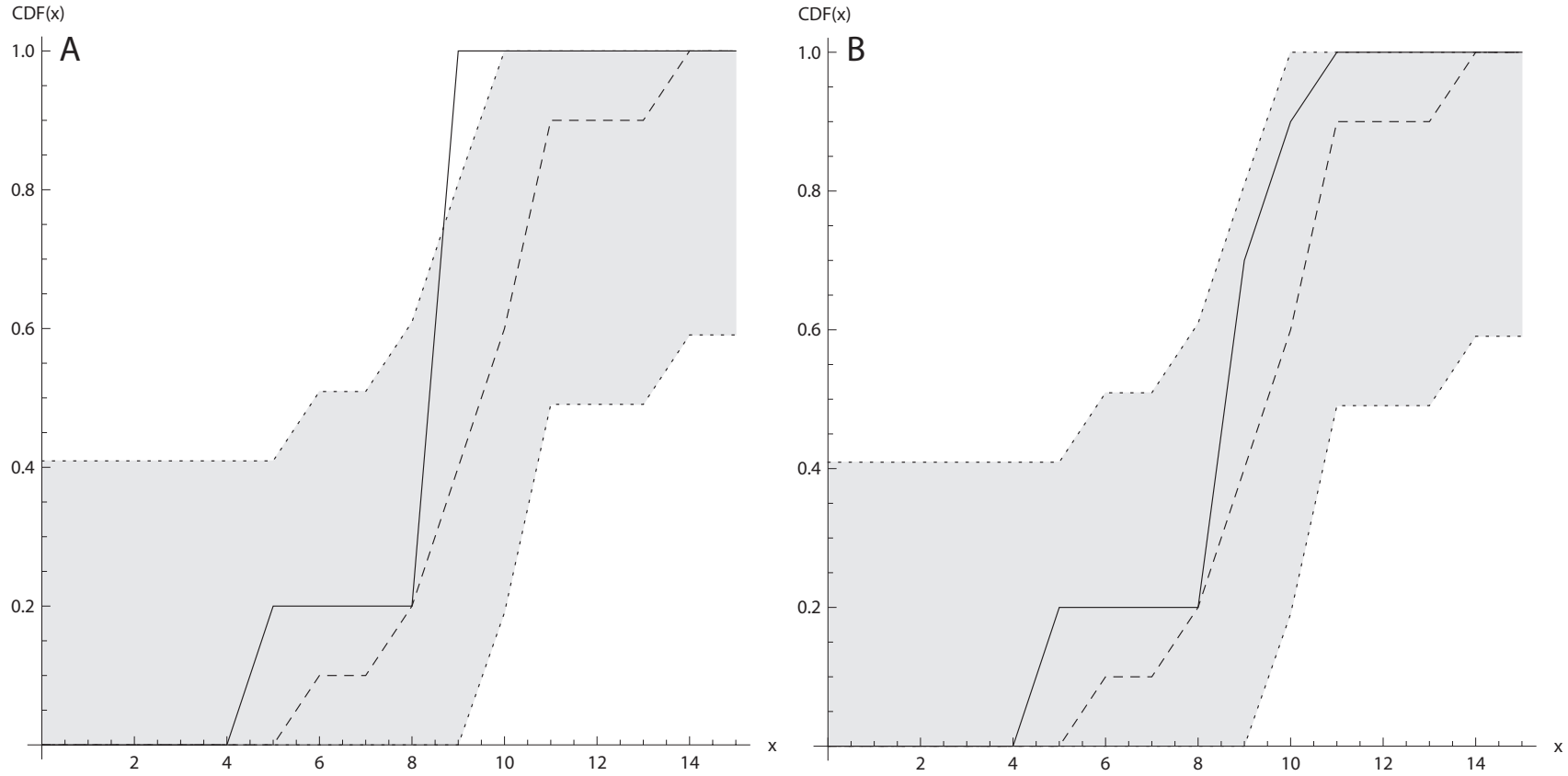
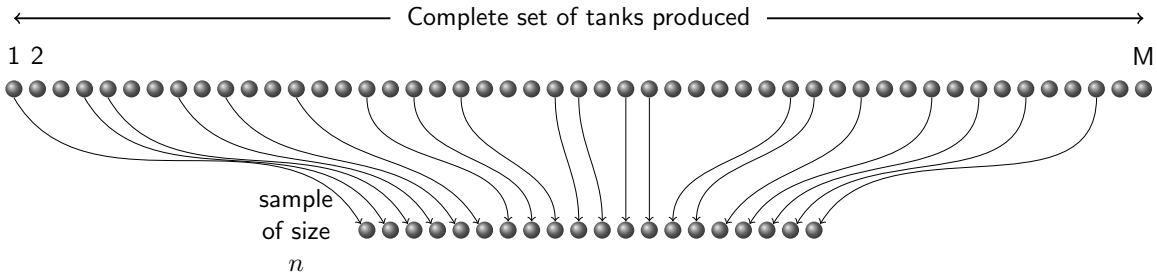


Figure: Empirical CDFs of (A) an infeasible and of (B) a feasible set of random variates O_2; these are $\{5, 5, 9, 9, 9, 9, 9, 9, 9, 9\}$ and $\{5, 5, 9, 9, 9, 9, 9, 10, 10, 11\}$, respectively.

Applications

German tank problem

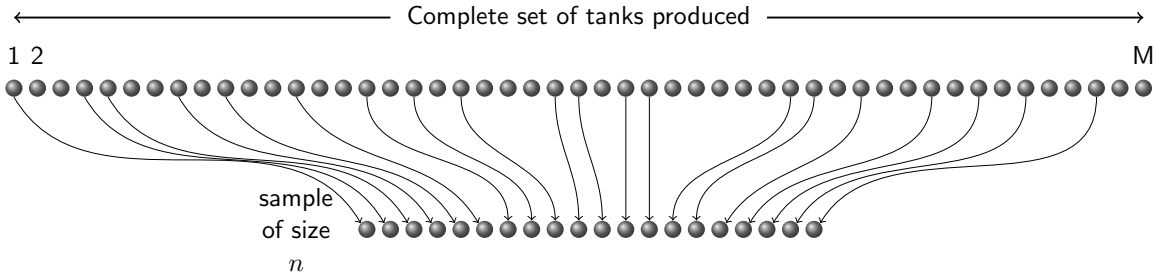
During World War II, production of German tanks such as the Panther was accurately estimated by Allied intelligence using statistical methods.



Estimate M by using information from captured tanks.

Applications

German tank problem

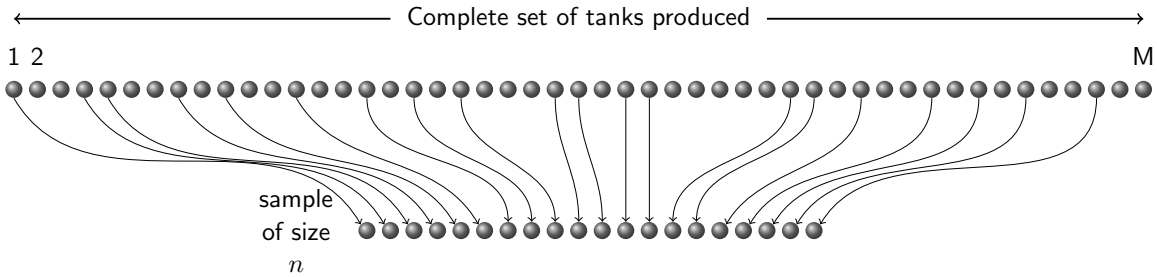


A sample $\{\omega_1, \ldots, \omega_n\}$ follows a multivariate hypergeometric distribution (urn model without replacement and with multiple classes of objects).

However, we will consider an approximation that exploits a Uniform$(0, M)$ distribution.

Applications

German tank problem



Consider a sample $\{\omega_1, \ldots, \omega_n\}$ from Uniform(0,M). Let $\omega'_{\text{max}} = \max\{\omega'_1, \ldots, \omega'_n\}$.

The test statistic ω_{max} follows the distribution

$$F(x) = \underbrace{\frac{x}{M} \cdots \frac{x}{M}}_{n \text{ times}} = \frac{x^n}{M^n} \qquad f(x) = n\frac{x^{n-1}}{M^n}$$

The confidence interval for the estimated population maximum is $(\omega'_{\text{max}}, \omega'_{\text{max}}/\alpha^{1/n})$, where $1 - \alpha$ is the confidence level sought.

Applications

German tank problem

In the following CSP, random variates have been generated from a Uniform(0,20).

Constraints:

(1) KS-test$^{\alpha}_{=}(O_1, \text{Uniform}(0, m))$

Decision variables:

$o_1 \in \{2\}, o_2 \in \{6\}, o_3 \in \{6\}, o_4 \in \{17\}, o_5 \in \{4\},$
$o_6 \in \{11\}, o_7 \in \{10\}, o_8 \in \{7\}, o_9 \in \{2\}, o_{10} \in \{15\},$
$m \in \{0, \infty\}$
$O_1 \equiv \{o_1, \ldots, o_{10}\}$
$\alpha = 0.05$

Figure: A CSP formulation of the German tank problem

After propagating constraint (1), the domain of m reduces to $\{9, \ldots, 28\}$, so at significance level $\alpha = 0.05$ we reject the null hypothesis that the true maximum M is outside this range.

Applications

German tank problem

In the following CSP, random variates have been generated from a Uniform(0,20).

Constraints:

(1) KS-test$^{\alpha}_{=}(O_1, \text{Uniform}(0, m))$

Decision variables:

$o_1 \in \{2\}, o_2 \in \{6\}, o_3 \in \{6\}, o_4 \in \{17\}, o_5 \in \{4\},$
$o_6 \in \{11\}, o_7 \in \{10\}, o_8 \in \{7\}, o_9 \in \{2\}, o_{10} \in \{15\},$
$m \in \{0, \infty\}$
$O_1 \equiv \{o_1, \ldots, o_{10}\}$
$\alpha = 0.05$

Figure: A CSP formulation of the German tank problem

Note that the parametric statistical approach previously discussed would produce a tighter interval: (17.0, 22.93).

Applications

Incomplete German tank problem

Now assume that soldiers have erased the last figure of the 6th tank serial number.

Constraints:

(1) KS-test$^{\alpha}_{=}(O_1, \text{Uniform}(0, m))$

Decision variables:

$o_1 \in \{2\}, o_2 \in \{6\}, o_3 \in \{6\}, o_4 \in \{17\}, o_5 \in \{4\},$
$o_6 \in \{10, \ldots, 19\}, o_7 \in \{10\}, o_8 \in \{7\}, o_9 \in \{2\},$
$o_{10} \in \{15\},$
$m \in \{0, \infty\} \quad O_1 \equiv \{o_1, \ldots, o_{10}\} \quad \alpha = 0.05$

Figure: A CSP formulation of the German tank problem

After propagating constraint (1), the domain of m reduces to $\{9, \ldots, 32\}$, so at significance level $\alpha = 0.05$ we reject the null hypothesis that the true maximum is outside this range.

Applications

Inspection scheduling

Parameters:

$U = 10$	Units to be inspected
$I = 25$	Inspections per unit
$H = 365$	Periods in the planning horizon
$D = 1$	Duration of an inspection
$M = 36$	Max interval between two inspections
$C = 1$	Inspectors required for an inspection
$m = 5$	Inspectors available
$\lambda = 1/5$	Inspection rate

Constraints:

(1) $\text{cumulative}(s, e, t, c, m)$

for all $u \in 1, \ldots, U$

(2) $\text{KS-test}^{\alpha}_{=}(O_u, \text{exponential}(\lambda))$

(3) $e_{uI} \geq H - M$

for all $u \in 1, \ldots, U$ **and** $j \in 2, \ldots, I$

(4) $i_{u,j-1} = s_{uI+j} - s_{uI+j-1} - 1$

(5) $s_{uI+j} \geq s_{uI+j-1}$

Decision variables:

$s_k \in \{1, \ldots, H\},$	$\forall k \in 1, \ldots, I \cdot U$
$e_k \in \{1, \ldots, H\},$	$\forall k \in 1, \ldots, I \cdot U$
$t_k \leftarrow D,$	$\forall k \in 1, \ldots, I \cdot U$
$c_k \leftarrow C,$	$\forall k \in 1, \ldots, I \cdot U$
$i_{u,j-1} \in \{0, \ldots, M\},$	$\forall u \in 1, \ldots, U$ and $\forall j \in 2, \ldots, I$
$O_u \equiv \{i_{u,1}, \ldots, i_{u,I-1}\},$	$\forall u \in 1, \ldots, U$

Figure: Inspection scheduling

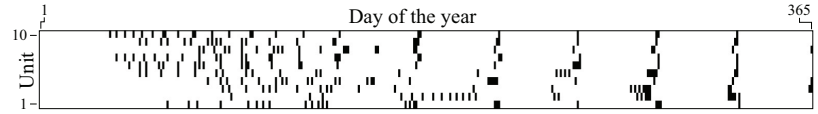


Figure: Inspection plan; black marks denote inspections.

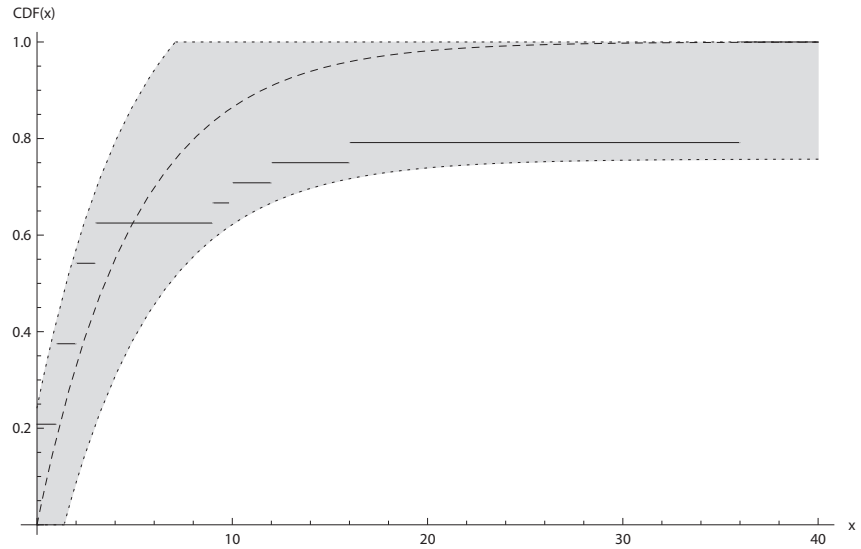
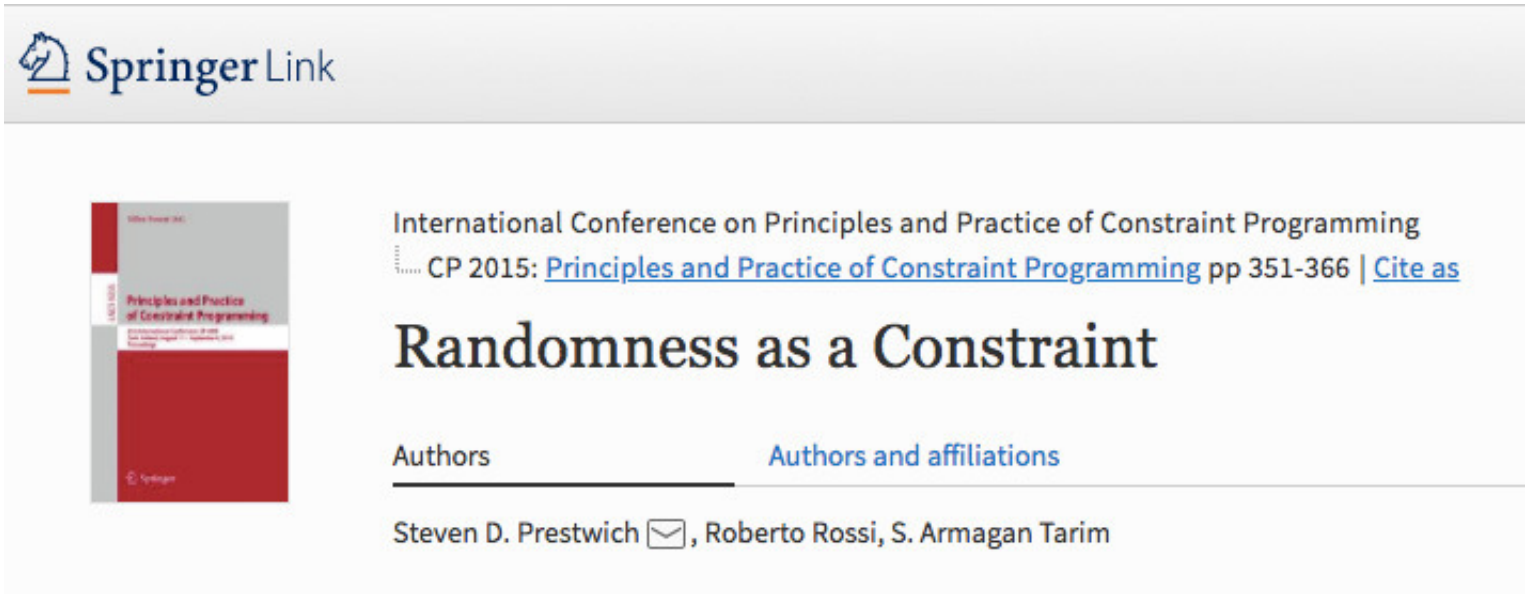


Figure: Empirical CDF of intervals (in days) between inspections for unit of assessment 1

Related works

Randomness as a constraint



S. Prestwich, R. Rossi and S. A. Tarim "Randomness as a Constraint", *in Proceedings of the 21st International Conference on Principles and Practice of Constraint Programming* (CP-2015), August 31 - September 4, 2015, Cork, Ireland, Lecture Notes in Computer Science, Springer-Verlag, LNCS 9255, pp.351-366, 2015

Related works

Declarative statistics

R. Rossi, O. Agkun, S. Prestwich, A. Tarim, "Declarative Statistics," University of Edinburgh, technical report, 2017

Statistical Programming

What is it?

Statistical Programming provides a platform for modeling a family of hypotheses about a phenomenon being studied and for determining which of these hypotheses is "compatible" with available data.

In a **Statistical Program**...

A feasible assignment represents one of the many possible hypotheses that are "compatible" with available data, i.e. that — to use statistical terminology — we failed to reject.

An infeasible assignments represent hypotheses that have been rejected, in a statistical sense, at the prescribed significance level on the basis of the observations that are available.

Statistical Programming

What can it do for me?

Statistical Programming can be used to model complex families of statistical hypotheses and query associated confidence regions.

These regions can be **queried** in different ways:

- the decision maker may ask what vector in the region is the *most likely one*, or
- *explore the boundaries* of the region to construct confidence intervals for model parameters.

Statistical Programming

Supporting technologies

`http://www.ibex-lib.org/`
`http://www.choco-solver.org/`

The t-test statistical constraint.

The χ^2 goodness of fit statistical constraint.

The χ^2 test of independence statistical constraint.

Fisher's ratio statistical constraint.

Hotelling's statistical constraints (Hotelling χ^2 or t^2).

Statistical Programming

Example: Linear model fitting

Objective:

$\min\ s$

Constraints:

(1)	$e_t = v_t - (at + b)$	for $t = 1, \ldots, T$
(2)	$\tau_i = T(F_\sigma(w_{i+1}) - F_\sigma(w_i))$	for $i = 1, \ldots, m$
(3)	$\chi^2_w(r; \tau; s)$	
(4)	$s \leq Q$	

Parameters:

T	time periods
$v_1, \ldots, v_T$	random variates
m	number of bins
$w_1, \ldots, w_{m+1}$	bin boundaries
F_t	normal cumulative distribution with mean 0 and standard deviation σ
Q	$1 - \alpha$ quantile of the inverse χ^2 distribution with $m - 1$ degrees of freedom

Decision variables:

a	fitted model slope
b	fitted model intercept
σ	normal standard deviation
$e_1, \ldots, e_T$	errors[1]
$\tau_1, \ldots, \tau_m$	target counts for each bin
s	χ^2 statistics

Figure: Statistical Programming model for linear model fitting.

[1] Note that these are errors, not residuals. Recall that a statistical error (or disturbance) is the amount by which an observation differs from its *unobservable* expected value.

Statistical Programming

Example: Linear model fitting

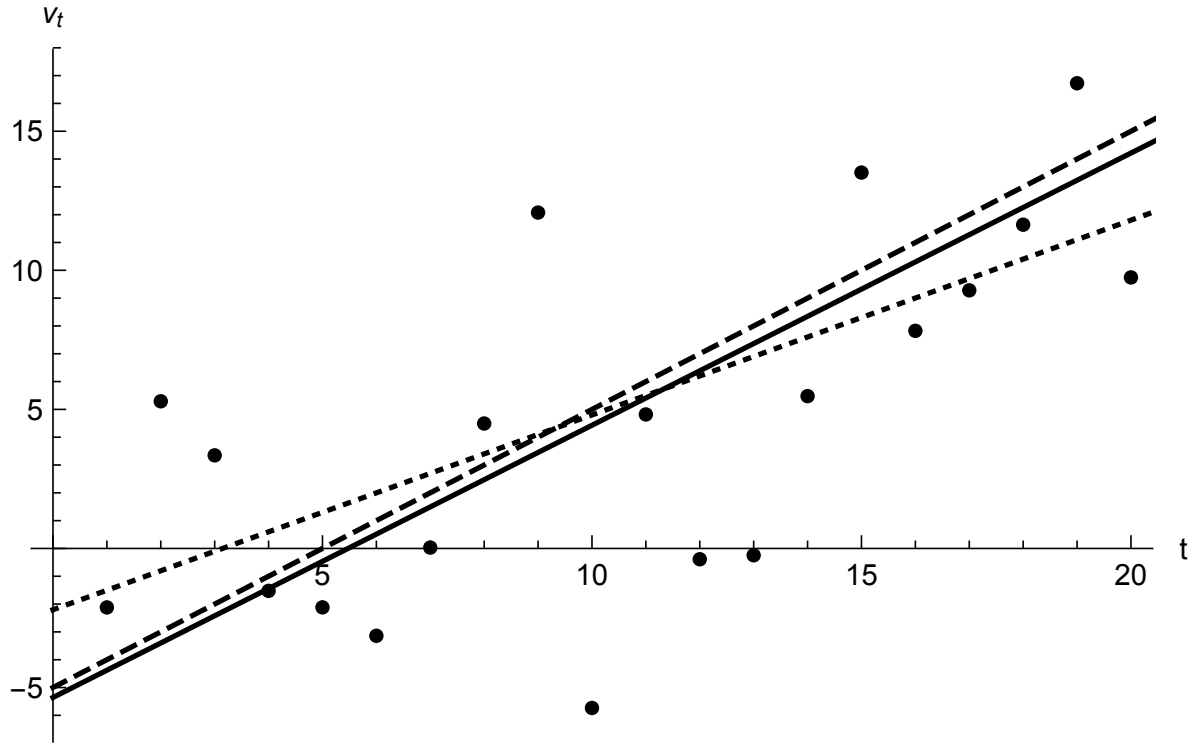


Figure: Linear model fitting. The dots represent random variates; the solid line is the model that generated these variates, which we are trying to estimate. The dotted line, is the model obtained via the method of least squares. The dashed line is the model obtained via Statistical Programming, by solving the Statistical Programming model.

Statistical Programming

Example: Linear model fitting

Constraints:

(1) $e_t = v_t - (at + b)$ for $t = 1, \ldots, T$

(2) $\chi^2_\alpha(e; \mu; \sigma)$

Parameters:

T	time periods
$v_1, \ldots, v_T$	random variates
μ	normal mean, fixed and set to 0
σ	normal standard deviation

Decision variables:

a	fitted model slope
b	fitted model intercept
$e_1, \ldots, e_T$	errors

Figure: Statistical Programming model for linear model fitting: known σ.

Statistical Programming

Example: Linear model fitting

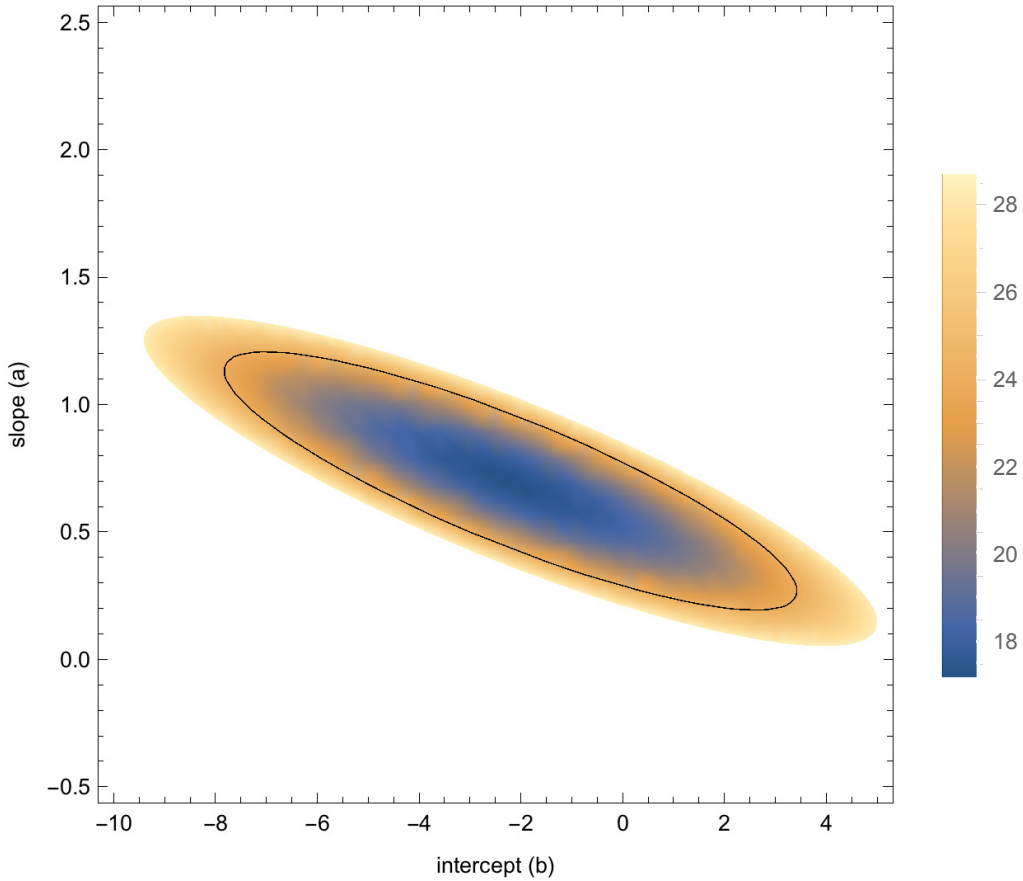


Figure: Confidence region for slope (a) and intercept (b) under the assumption that σ is known; the colour gradient reflects the value of the χ^2 statistics.

Statistical Programming

Example: Linear model fitting

Constraints:

(1)	$e_t = v_t - (at + b)$	for $t = 1, \ldots, T$
(2)	$t^2_\alpha(e; \mu)$	

Parameters:

T	time periods
$v_1, \ldots, v_T$	random variates

Decision variables:

a	fitted model slope
b	fitted model intercept
$e_1, \ldots, e_T$	errors

Figure: Statistical Programming model for linear model fitting: unknown σ.

Statistical Programming

Example: Linear model fitting

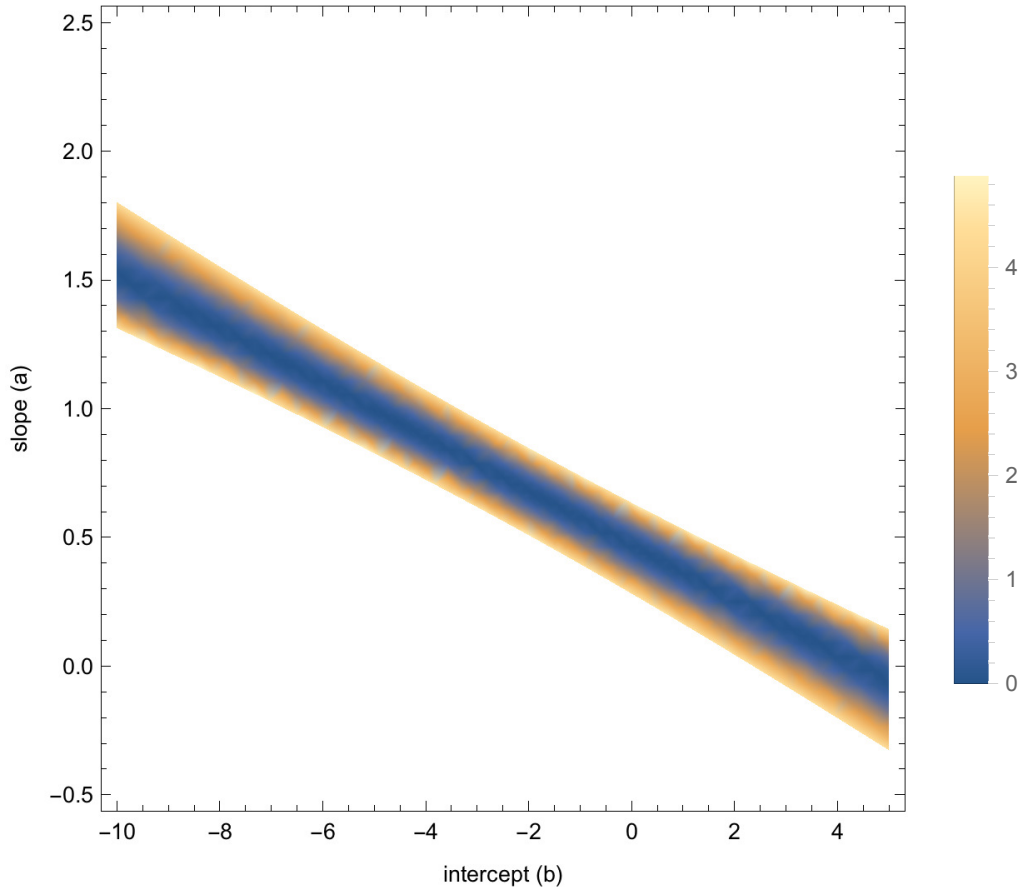


Figure: Confidence region for slope (a) and intercept (b) under the assumption that σ is estimated from the data; the colour gradient reflects the value of the t^2 statistics.

Statistical Programming

Example: ANOVA

Group 1	Group 2	Group 3
3.57329	9.83132	9.80335
6.5655	9.7379	8.79726
-2.06033	6.6339	13.6045
0.469477	8.20049	9.4932
3.05632	7.19737	8.50685
5.54063	9.19586	9.22433

Table: Samples considered in our example.

Statistical Programming

Example: ANOVA

		DF	SumOfSq	MeanSq	F-Ratio	p-value
ANOVA	Model	2	166.387	83.1935	16.0089	0.000189984
	Error	15	77.9505	5.1967		
	Total	17	244.337			
Means	Overall	7.07618				
	Group 1	2.85748				
	Group 2	8.46614				
	Group 3	9.90492				

Table: One-way analysis of variance table; DF: degrees of freedom; SumOfSq: sum of squared differences; MeanSq: mean squared differences.

In the case of our numerical example a post-hoc Tukey's test would reveal that the mean of Group 1 differ at the prescribed significance level from that of Group 2 and Group 3.

Statistical Programming

Example: ANOVA as a Statistical Programming model

Constraints:

"between group" mean squared differences

(1) $\text{MEAN}(\bar{y}_i, \{O_{i,1}, \ldots, O_{i,n}\})$ for $i = 1, \ldots, m$

(2) $\text{VARIANCE}(s_b, \{\bar{y}_1, \ldots, \bar{y}_m\})$

(3) $\bar{s}_b = ns_b/(m-1)$

"within group" mean squared differences

(4) $\text{VARIANCE}(s_w^i, \{O_{i,1}, \ldots, O_{i,n}\})$ for $i = 1, \ldots, m$

(5) $\text{MEAN}(\bar{s}_w, \{s_w^1, \ldots, s_w^m\})$

Fisher's F-ratio statistic

(6) $\bar{s}_b/\bar{s}_w \leq F^{-1}_{m-1,m(n-1)}(1-\alpha)$

Parameters:

m	number of groups
n	number of random variates within a group
$O_{i,j}$	random variate j in group i
α	significance level

Decision variables:

$\bar{y}_i$	mean within group i
ns_b	"between group" sum of squared differences
$\bar{s}_b$	"between group" mean squared differences
$\bar{s}_w$	"within group" mean squared differences
$\bar{s}_b/\bar{s}_w$	F-ratio statistic
$F_{a,b}$	F-ratio distribution with a degrees of freedom in the in the numerator and b in the denominator.

Figure: Statistical Programming model for one-way analysis of variance.

Statistical Programming

Example: An alternative Statistical Programming model

Objective:

$\min\ s$

Constraints:

(1) $t^2(\langle O \rangle; \mu; s)$

(2) $s \leq Q$

Parameters:

m	number of groups
n	number of random variates within a group
$O_{i,j}$	random variate j in group i
Q	$1-\alpha$ quantile of the inverse Hotelling's $T^2_{m,n-1}$ distribution

Decision variables:

μ_i	mean of group i
s	Hotelling's t^2 statistic

Figure: Statistical Programming model for multivariate normal parameter fitting.

Statistical Programming

Example: An alternative Statistical Programming model

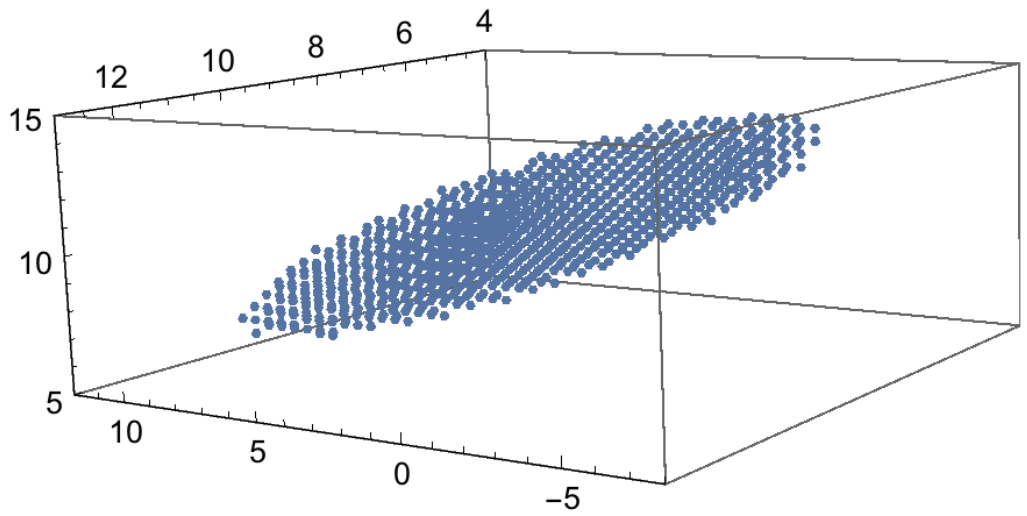


Figure: Statistical Programming model for multivariate normal parameter fitting: confidence region.

Statistical Programming

Example: An alternative Statistical Programming model

Constraints:

(1) $t^2(\langle O \rangle; \mu; s)$
(2) $s \leq Q$
(3) $\mu_i = \mu_j$ for all i, j where $i < j$

Parameters:

m	number of groups
n	number of random variates within a group
$O_{i,j}$	random variate j in group i
Q	$1 - \alpha$ quantile of the inverse Hotelling's $T^2_{m,n-1}$ distribution

Decision variables:

μ_i	mean of group i
s	Hotelling's t^2 statistic

Figure: Statistical Programming model for testing equality of means (infeasible).

Statistical Programming

Example: An alternative Statistical Programming model

Constraints:

(1) $t^2(\langle O \rangle; \mu; s)$
(2) $s \leq Q$
(3) $\mu_2 = \mu_3$

Parameters:

m	number of groups
n	number of random variates within a group
$O_{i,j}$	random variate j in group i
Q	$1 - \alpha$ quantile of the inverse Hotelling's $T^2_{m,n-1}$ distribution

Decision variables:

μ_i	mean of group i
s	Hotelling's t^2 statistic

Figure: Statistical Programming model for testing equality of means (feasible... in line with Tukey's test!).

Statistical Programming

Example: An alternative Statistical Programming model

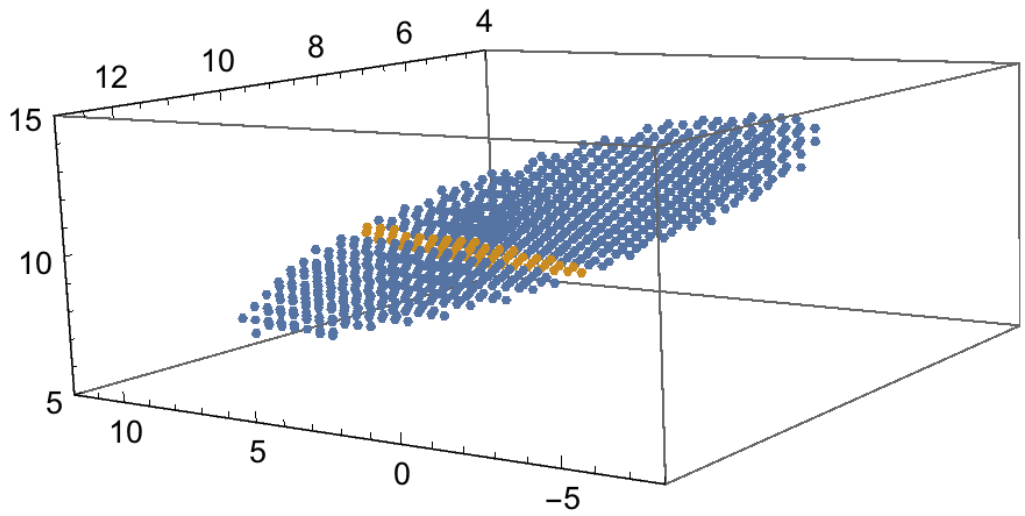


Figure: Statistical Programming model for multivariate normal parameter fitting: reduced confidence region under $\mu_2 = \mu_3$.

Statistical Programming

Example: An alternative Statistical Programming model

Objective:

$\min$ (or $\max$) μ_i

Constraints:

(1) $t^2(\langle O \rangle; \mu; s)$

(2) $s \leq Q$

Parameters:

m	number of groups
n	number of random variates within a group
$O_{i,j}$	random variate j in group i
Q	$1 - \alpha$ quantile of the inverse Hotelling's $T^2_{m,n-1}$ distribution

Decision variables:

μ_i	mean of group i
s	Hotelling's t^2 statistic

Figure: Statistical Programming model for deriving μ_i confidence interval.

Statistical Programming

Example: An alternative Statistical Programming model

Objective:

$\min$ (or $\max$) μ_i

Constraints:

(1) $t^2(\langle O \rangle; \mu; s)$
(2) $s \leq Q$
(3) $\mu_2 = \mu_3$

Parameters:

m	number of groups
n	number of random variates within a group
$O_{i,j}$	random variate j in group i
Q	$1 - \alpha$ quantile of the inverse Hotelling's $T^2_{m,n-1}$ distribution

Decision variables:

μ_i	mean of group i
s	Hotelling's t^2 statistic

Figure: Statistical Programming model for deriving μ_i confidence interval under the assumption that $\mu_2 = \mu_3$.

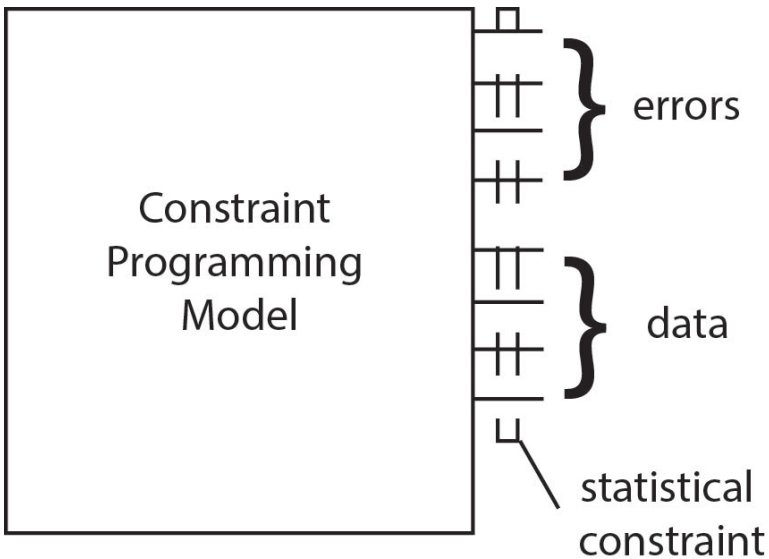
Statistical Programming

Applications discussed in the technical report

Application	*Statistical constraint(s)*
Linear model fitting	χ^2 goodness of fit
Time series analysis	χ^2 goodness of fit
	χ^2 test of independence
ANOVA	Fisher's ratio
Comparing means of two or more samples	Hotelling's t^2
multinomial proportions conf. interv.	Hotelling's χ^2 and t^2

Table: Applications considered and associated statistical constraints.

Statistical Programming



Constraints:

(1) $\mu_i = \mu_j$ for all i, j where $i < j$

(2) $t^2(\langle O \rangle; \mu; \alpha)$ ←statistical constraint

Parameters:

m	number of groups
n	number of random variates within a group
$O_{i,j}$	random variate j in group i
α	significance level

Decision variables:

μ_i	mean of group i

Statistical Programming

View on GitHub

syat-choco

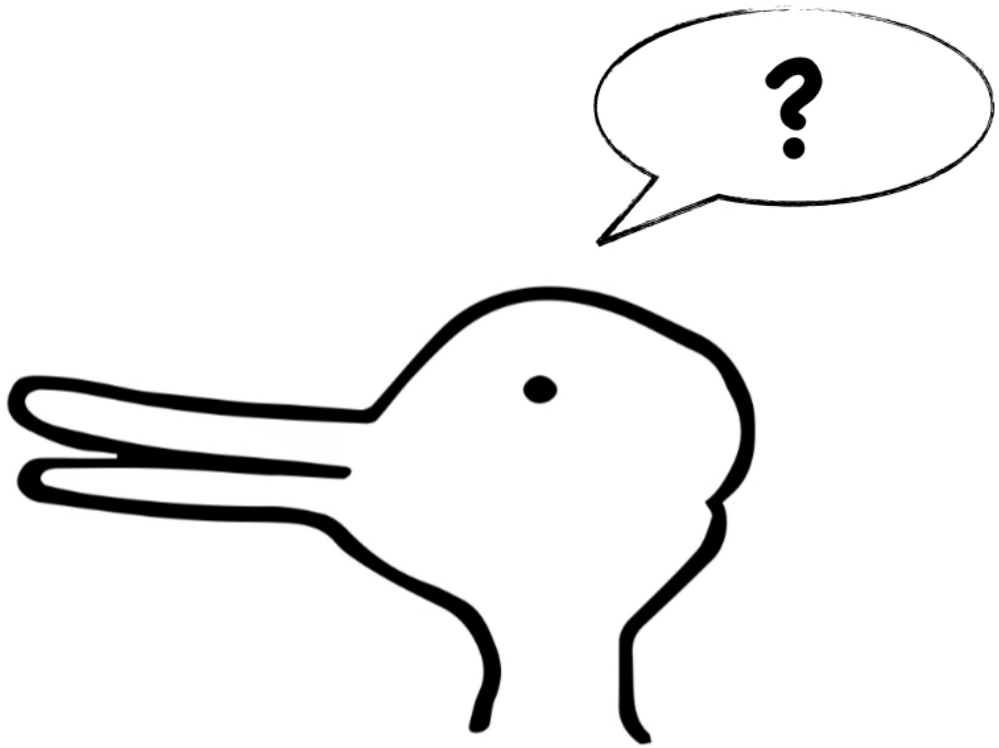
a Choco extension for Declarative Statistics

syat-choco is a Choco extension for Declarative Statistics.

Syādvāda, in Jaina metaphysics, the doctrine that all judgments are conditional, holding good only in certain conditions, circumstances, or senses, expressed by the word syāt (Sanskrit: "may be"). The ways of looking at a thing (called naya) are infinite in number.

For more information, please consult our Wiki.

Maven release:

```
<groupId>com.github.gwr3n</groupId>
<artifactId>syat-choco</artifactId>
<version>1.0.0</version>
```

Questions