Statistics and learning

An introduction: from data to modelling

Emmanuel Rachelson and Matthieu Vignes

ISAE SupAero

Wednesday 3rd September 2013

Statistical approach

A quick, partial and not very comprehensive overview

- Goal of this course: not a recipe cooking handbook, rather a path to **mathematical reasoning** which leads to dealing with **quantitative aspects of decision making** from data and still accounting for **uncertainty**.

Statistical approach

A quick, partial and not very comprehensive overview

- Goal of this course: not a recipe cooking handbook, rather a path to **mathematical reasoning** which leads to dealing with **quantitative aspects of decision making** from data and still accounting for **uncertainty**.
- Application of stats (and Machine Learning): pension matters, optimal insurance premium, stock market predictions, clinical trial for a new medicine, marketing research and planning ... Different branches: geostatistics, statistical physics ...

Statistical approach

A quick, partial and not very comprehensive overview

- Goal of this course: not a recipe cooking handbook, rather a path to **mathematical reasoning** which leads to dealing with **quantitative aspects of decision making** from data and still accounting for **uncertainty**.
- Application of stats (and Machine Learning): pension matters, optimal insurance premium, stock market predictions, clinical trial for a new medicine, marketing research and planning . . . Different branches: geostatistics, statistical physics . . .
- **Professional** AND **citizen interest**: ad-hoc exploitation of available data and don't be manipulated ?!

Statistical approach

A quick, partial and not very comprehensive overview

- Goal of this course: not a recipe cooking handbook, rather a path to **mathematical reasoning** which leads to dealing with **quantitative aspects of decision making** from data and still accounting for **uncertainty**.
- Application of stats (and Machine Learning): pension matters, optimal insurance premium, stock market predictions, clinical trial for a new medicine, marketing research and planning . . . Different branches: geostatistics, statistical physics . . .
- **Professional** AND **citizen interest**: ad-hoc exploitation of available data and don't be manipulated ?!
- few prerequisites: basic/intermediate maths and probability calculus.

Statistical approach

A quick, partial and not very comprehensive overview

- Goal of this course: not a recipe cooking handbook, rather a path to **mathematical reasoning** which leads to dealing with **quantitative aspects of decision making** from data and still accounting for **uncertainty**.
- Application of stats (and Machine Learning): pension matters, optimal insurance premium, stock market predictions, clinical trial for a new medicine, marketing research and planning ... Different branches: geostatistics, statistical physics ...
- **Professional** AND **citizen interest**: ad-hoc exploitation of available data and don't be manipulated ?!
- few prerequisites: basic/intermediate maths and probability calculus.
- Grail: linking data to mathematical modelling, objectivelly quantify and interpret conclusions and...awareness of limitations: **statistics helps but won't make decision for you** !

Inspiring work / our bibliography

T. Hastie, R. Tibshirani and J. Friedman.
Elements of statistical learning.
Springer, 2nd edition, 2009.

E. Moulines, F. Roueff and J.-L. Pac (and formerly F. Rossi)
Statistiques.
Cours TelecomParisTech, 2008.

A. Garivier
Statistiques avancées.
Cours Centrale 2011, 2011.

S. Clémençon.
Apprentissage statistique.
Cours TELECOM ParisTech, 2011-2012.

S. Arlot, Francis B., O. Catoni, G. Stolz and G. Obozinski
Apprentissage.
Cours ENS, 2012.

N. Chopin, D. Rosenberg and G. Stolz
Eléments de statistique pour citoyens d'aujourd'hui et managers de demain.
Cours L3 HEC, 2012–2013.

A. Baccini, P. Besse, S. Canu, S. Déjean, B. Laurent, C. Marteau, P. Martin and H. Milhem
Wikistat, le cours dont vous êtes le héros.
`http://wikistat.fr/`, 2012.

A.B. Dufour D. Chessel J.R. Lobry, S. Mousset and S. Dray
Enseignements de Statistique en Biologie.
`http://pbil.univ-lyon1.fr/R/`, 2012.

F. Bertrand
Page professionnelle - Enseignements.
`http://www-irma.u-strasbg.fr/~fbertran/enseignement/`, 2012.

V. Montbet
Page professionnelle - Enseignements.
`http://perso.univ-rennes1.fr/valerie.monbet/enseignement.html`, 2013.

And many others we just forgot to mention.

From data to modelling
and back

Two different situations might occur for the same modelling:

- empirical approach to gaining knowledge from an experiment repeated many times,

From data to modelling

and back

Two different situations might occur for the same modelling:

- empirical approach to gaining knowledge from an experiment repeated many times,
- study of a sample drawn from a population.

From data to modelling

and back

Two different situations might occur for the same modelling:

- empirical approach to gaining knowledge from an experiment repeated many times,
- study of a sample drawn from a population.

Preference between two possible configurations

Consumer ID	1	2	3	4	5	6	...
Opinion	A	A	B	A	B	B	...

We can denote by x_i successive opinions taking (binary) values "A" $(=0)$ or "B" $(=1)$. Mathematician sees that as realisation of random variables denoted X_i.

Localising randomness

Randomness...

...arises from the choice of the questioned persons, NOT from in each actual answer.

Incidental reminder: Bernouilli distribution, with parameter $0 < p < 1$...

- laid question: is $p_0 > 1/2$ or $< 1/2$? This is a **test**.

Localising randomness

Randomness...

...arises from the choice of the questioned persons, NOT from in each actual answer.

Incidental reminder: Bernouilli distribution, with parameter $0 < p < 1$...

- laid question: is $p_0 > 1/2$ or $< 1/2$? This is a **test**.
- but hey, what is p_0 actually ??

Localising randomness

Randomness...

...arises from the choice of the questioned persons, NOT from in each actual answer.

Incidental reminder: Bernouilli distribution, with parameter $0 < p < 1$...

- laid question: is $p_0 > 1/2$ or $< 1/2$? This is a **test**.
- but hey, what is p_0 actually ??
- all consumers cannot be interviewed for obvious reasons !

Localising randomness

Randomness...

...arises from the choice of the questioned persons, NOT from in each actual answer.

Incidental reminder: Bernouilli distribution, with parameter $0 < p < 1$...

- laid question: is $p_0 > 1/2$ or $< 1/2$? This is a **test**.
- but hey, what is p_0 actually ??
- all consumers cannot be interviewed for obvious reasons !
- Statistics is a sound framework to

Localising randomness

Randomness...

...arises from the choice of the questioned persons, NOT from in each actual answer.

Incidental reminder: Bernouilli distribution, with parameter $0 < p < 1$...

- laid question: is $p_0 > 1/2$ or $< 1/2$? This is a **test**.
- but hey, what is p_0 actually ??
- all consumers cannot be interviewed for obvious reasons !
- Statistics is a sound framework to
    1. describe sample using estimates

Localising randomness

Randomness...

...arises from the choice of the questioned persons, NOT from in each actual answer.

Incidental reminder: Bernouilli distribution, with parameter $0 < p < 1$...

- laid question: is $p_0 > 1/2$ or $< 1/2$? This is a **test**.
- but hey, what is p_0 actually ??
- all consumers cannot be interviewed for obvious reasons !
- Statistics is a sound framework to
  1. describe sample using estimates
  2. quantitatively answer the question (generalising sample to full population conclusions)

Modelling

a mandatory step

- we start by modelling the problem (Sample/population description, variables and their distributions).

Modelling

a mandatory step

- we start by modelling the problem (Sample/population description, variables and their distributions).
- in the example: "correct model" $\in (\mathrm{Bern}(p))_p$; n realisations $(x_i)_i$ of iid (indep. & identical. distr.) random variables $(X_i)_i \sim \mathrm{Bern}\,(p)$ are available.

Modelling

a mandatory step

- we start by modelling the problem (Sample/population description, variables and their distributions).
- in the example: "correct model" $\in (\mathrm{Bern}(p))_p$; n realisations $(x_i)_i$ of iid (indep. & identical. distr.) random variables $(X_i)_i \sim \mathrm{Bern}\,(p)$ are available.
- remember the Jean Tibéri vs. Lyne Cohen-Solal (+ Ph. Meyer) council election in Paris in 2008 between 20.45 and 21.15 ? At 20.45, $(463; 409; 106)$ but after counting the votes :
  $(11,044; 11,269; 2,730)$.

Modelling

a mandatory step

- we start by modelling the problem (Sample/population description, variables and their distributions).
- in the example: "correct model" $\in (\text{Bern}(p))_p$; n realisations $(x_i)_i$ of iid (indep. & identical. distr.) random variables $(X_i)_i \sim \text{Bern}(p)$ are available.
- remember the Jean Tibéri vs. Lyne Cohen-Solal (+ Ph. Meyer) council election in Paris in 2008 between 20.45 and 21.15 ? At 20.45, $(463; 409; 106)$ but after counting the votes : $(11,044; 11,269; 2,730)$.
- Construction of **confidence intervals** to answer the question.

Useful tools

a reminder ?

- **mean** (central tendency): $E[X] = \bar{X} := \frac{X_1 + ... + X_n}{n}$,

Useful tools

a reminder ?

- **mean** (central tendency): $E[X] = \bar{X} := \frac{X_1 + \ldots + X_n}{n}$,
- **standard deviation** (dispersion tendency): $= \sigma(X) := \sqrt{E\left[(X - E[X])^2\right]} = \sqrt{E[X]^2 - E[X^2]}$,

Useful tools

a reminder ?

- **mean** (central tendency): $E[X] = \bar{X} := \frac{X_1 + \ldots + X_n}{n}$,
- **standard deviation** (dispersion tendency):
  $= \sigma(X) := \sqrt{E\left[(X - E[X])^2\right]} = \sqrt{E[X]^2 - E[X^2]}$,
- (almost never use skewness and kurtosis)

Two important probabilistic tools in statistics

Law of large numbers

Theorem

Let $X_1 \dots X_n$ be iid random variables with mean μ. Then the empirical mean converges in probability towards μ, i.e.:

$$\overline{X_n} := \frac{1}{n}(X_1 + \ldots + X_n) \longrightarrow \mu.$$

Two important probabilistic tools in statistics

Law of large numbers

Theorem

Let $X_1 \dots X_n$ be iid random variables with mean μ. Then the empirical mean converges in probability towards μ, i.e.:

$$\overline{X_n} := \frac{1}{n}(X_1 + \ldots + X_n) \longrightarrow \mu.$$

In other term, for all $\epsilon > 0$, $P\left(\mid \overline{X_n} - \mu \mid > \epsilon\right) \to 0$

Two important probabilistic tools in statistics

Central limit theorem

Theorem

Let $X_1 \dots X_n$ be iid random variables which admit an order 2 moment. Denote by μ and σ the corresponding mean and standard deviation, then:

$$\frac{\sqrt{n}}{\sigma}(\overline{X_n} - \mu) \longrightarrow \mathcal{N}(0,1).$$

Two important probabilistic tools in statistics

Central limit theorem

Theorem

Let $X_1 \dots X_n$ be iid random variables which admit an order 2 moment. Denote by μ and σ the corresponding mean and standard deviation, then:

$$\frac{\sqrt{n}}{\sigma}(\overline{X_n} - \mu) \longrightarrow \mathcal{N}(0,1).$$

In the case of distribution with density functions, this means that

$$P\left(\frac{\sqrt{n}}{\sigma}(\overline{X_n} - \mu) \leq x\right) := F_n(x) \longrightarrow P(Z \leq x) = \frac{\int_{-\infty}^{x} \mathrm{e}^{-z^2/2}\, dz}{\sqrt{2\pi}}.$$

Deciding from a sample

- Back to the "preference" example: A vs. B.

Deciding from a sample

- Back to the "preference" example: A vs. B.

Preference between two possible configurations

Consumer ID	1	2	3	4	5	6	...
Opinion	A	A	B	A	B	B	...

Deciding from a sample

- Back to the "preference" example: A vs. B.

Preference between two possible configurations

Consumer ID	1	2	3	4	5	6	...
Opinion	A	A	B	A	B	B	...

- We compute $\overline{x_{100}} = \frac{x_1+\cdots+x_{100}}{100} = 0.42$.

Deciding from a sample

- Back to the "preference" example: A vs. B.

Preference between two possible configurations

Consumer ID	1	2	3	4	5	6	...
Opinion	A	A	B	A	B	B	...

- We compute $\overline{x_{100}} = \frac{x_1+\cdots+x_{100}}{100} = 0.42$.
- Our intuition and the LLN tell us that p_0 is "close" to 0.42.

Deciding from a sample

- Back to the "preference" example: A vs. B.

Preference between two possible configurations

Consumer ID	1	2	3	4	5	6	...
Opinion	A	A	B	A	B	B	...

- We compute $\overline{x_{100}} = \frac{x_1 + \cdots + x_{100}}{100} = 0.42$.
- Our intuition and the LLN tell us that p_0 is "close" to 0.42.
- Can we conclude ? Is this **estimate** enough ?

Let's play around the Central limit theorem...

Concluding the example

at the price of a slight risk

- we directly derive $\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}(\overline{X_n} - p_0) \rightarrow \mathcal{N}(0,1)$ so

Concluding the example

at the price of a slight risk

- we directly derive $\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}(\overline{X_n} - p_0) \rightarrow \mathcal{N}(0,1)$ so
- with a **confidence level** (what is that ??) of 95%: $\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}|\overline{X_n} - p_0| \leq u = 1.96$. This is equivalent to:

Concluding the example

at the price of a slight risk

- we directly derive $\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}(\overline{X_n} - p_0) \to \mathcal{N}(0,1)$ so
- with a **confidence level** (what is that ??) of 95%: $\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}|\overline{X_n} - p_0| \leq u = 1.96$. This is equivalent to:
- $p_0 \in I_n := \left[\overline{X_n} - 1.96\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}; \overline{X_n} + 1.96\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}\right],$

Concluding the example

at the price of a slight risk

- we directly derive $\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}(\overline{X_n} - p_0) \to \mathcal{N}(0,1)$ so
- with a **confidence level** (what is that ??) of 95%: $\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}|\overline{X_n} - p_0| \leq u = 1.96$. This is equivalent to:
- $p_0 \in I_n := \left[\overline{X_n} - 1.96\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}; \overline{X_n} + 1.96\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}\right]$,
- again, does this help ?

Concluding the example

at the price of a slight risk

- we directly derive $\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}(\overline{X_n} - p_0) \to \mathcal{N}(0,1)$ so
- with a **confidence level** (what is that ??) of 95%: $\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}|\overline{X_n} - p_0| \leq u = 1.96$. This is equivalent to:
- $p_0 \in I_n := \left[\overline{X_n} - 1.96\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}; \overline{X_n} + 1.96\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}\right]$,
- again, does this help ?
- yes, using the simplifying trick $\sqrt{x(1-x)} \leq 1/2$: $I_n \subseteq \left[\bar{X_n} - 1/\sqrt{n}; \bar{X_n} + 1/\sqrt{n}\right] = [32\%; 52\%]$ in our scenario. Your final conclusion ? Again what is 95% here ?

Concluding the example

at the price of a slight risk

- we directly derive $\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}(\overline{X_n} - p_0) \to \mathcal{N}(0,1)$ so
- with a **confidence level** (what is that ??) of 95%: $\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}|\overline{X_n} - p_0| \leq u = 1.96$. This is equivalent to:
- $p_0 \in I_n := \left[\overline{X_n} - 1.96\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}; \overline{X_n} + 1.96\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}\right]$,
- again, does this help ?
- yes, using the simplifying trick $\sqrt{x(1-x)} \leq 1/2$: $I_n \subseteq \left[\bar{X}_n - 1/\sqrt{n}; \bar{X}_n + 1/\sqrt{n}\right] = [32\%; 52\%]$ in our scenario. Your final conclusion ? Again what is 95% here ?
- is the conclusion similar if $n = 1,000$?

Note: 95% could have been replaced by 99%. How could this have affected the conclusion ? What about 100% ?

Hypothesis testing

decision depends on what is tested ?!

- (H_0) is the basic hypothesis. It will be rejected iif data strongly supports it (e.g. dangerous drug or alleged innocent). Then

Hypothesis testing

decision depends on what is tested ?!

- (H_0) is the basic hypothesis. It will be rejected iif data strongly supports it (e.g. dangerous drug or alleged innocent). Then
- (H_1) is the alternative hypothesis which would be accepted iif (H_0) was regognised to be unacceptable.

Hypothesis testing

decision depends on what is tested ?!

- (H_0) is the basic hypothesis. It will be rejected iif data strongly supports it (e.g. dangerous drug or alleged innocent). Then
- (H_1) is the alternative hypothesis which would be accepted iif (H_0) was regognised to be unacceptable.
- Back on the example: do you want to test (H_0) $p_0 \geq 1/2$? Or (H'_0) $p_0 \leq 1/2$? Which one is the sensible/aggressive choice ?

Hypothesis testing

decision depends on what is tested ?!

- (H_0) is the basic hypothesis. It will be rejected iif data strongly supports it (e.g. dangerous drug or alleged innocent). Then
- (H_1) is the alternative hypothesis which would be accepted iif (H_0) was regognised to be unacceptable.
- Back on the example: do you want to test (H_0) $p_0 \geq 1/2$? Or (H_0') $p_0 \leq 1/2$? Which one is the sensible/aggressive choice ?
- lessons from this: tests are not reductible to confidence intervals and...don't be fooled by an obscure choice of hypotheses !

Outline of the Statistics and learning course

From 18^{th} September 2013 to 7^{th} February 2014, you will hear about:

- descriptive statistics and modelling: estimation, dispersion measure, confidence intervals, PCA and perhaps more (CA, clustering, risk function).

- Projects with some stats content from Nov. 2013 'til May 2014...tbc

Outline of the Statistics and learning course

From 18^{th} September 2013 to 7^{th} February 2014, you will hear about:

- descriptive statistics and modelling: estimation, dispersion measure, confidence intervals, PCA and perhaps more (CA, clustering, risk function).
- Tests, ANOVA (one, several factors, analysis of covariance)
- Projects with some stats content from Nov. 2013 'til May 2014...tbc

Outline of the Statistics and learning course

From 18th September 2013 to 7th February 2014, you will hear about:

- descriptive statistics and modelling: estimation, dispersion measure, confidence intervals, PCA and perhaps more (CA, clustering, risk function).
- Tests, ANOVA (one, several factors, analysis of covariance)
- (linear) Regressions

- Projects with some stats content from Nov. 2013 'til May 2014...tbc

Outline of the Statistics and learning course

From 18^{th} September 2013 to 7^{th} February 2014, you will hear about:

- descriptive statistics and modelling: estimation, dispersion measure, confidence intervals, PCA and perhaps more (CA, clustering, risk function).
- Tests, ANOVA (one, several factors, analysis of covariance)
- (linear) Regressions
- Trees and ensemble methods

- Projects with some stats content from Nov. 2013 'til May 2014...tbc

Outline of the Statistics and learning course

From 18th September 2013 to 7th February 2014, you will hear about:

- descriptive statistics and modelling: estimation, dispersion measure, confidence intervals, PCA and perhaps more (CA, clustering, risk function).
- Tests, ANOVA (one, several factors, analysis of covariance)
- (linear) Regressions
- Trees and ensemble methods
- Kernels and parcimony

- Projects with some stats content from Nov. 2013 'til May 2014...tbc

Outline of the Statistics and learning course

From 18^{th} September 2013 to 7^{th} February 2014, you will hear about:

- descriptive statistics and modelling: estimation, dispersion measure, confidence intervals, PCA and perhaps more (CA, clustering, risk function).
- Tests, ANOVA (one, several factors, analysis of covariance)
- (linear) Regressions
- Trees and ensemble methods
- Kernels and parcimony
- Markov Chain Monte Carlo methods

- Projects with some stats content from Nov. 2013 'til May 2014...tbc

Outline of the Statistics and learning course

From 18^{th} September 2013 to 7^{th} February 2014, you will hear about:

- descriptive statistics and modelling: estimation, dispersion measure, confidence intervals, PCA and perhaps more (CA, clustering, risk function).
- Tests, ANOVA (one, several factors, analysis of covariance)
- (linear) Regressions
- Trees and ensemble methods
- Kernels and parcimony
- Markov Chain Monte Carlo methods
- ...and lots of R ;) !

- Projects with some stats content from Nov. 2013 'til May 2014...tbc