18.650
Statistics for Applications

Chapter 1: Introduction

Goals

Goals:

- To give you a solid introduction to the mathematical theory behind statistical methods;
- To provide theoretical guarantees for the statistical methods that you may use for certain applications.

At the end of this class, you will be able to

1. From a real-life situation, formulate a statistical problem in mathematical terms
2. Select appropriate statistical methods for your problem
3. Understand the implications and limitations of various methods

Instructors

- Instructor: Philippe Rigollet
  Associate Prof. of Applied Mathematics; IDSS; MIT Center for Statistics and Data Science.

- Teaching Assistant: Victor-Emmanuel Brunel
  Instructor in Applied Mathematics; IDSS; MIT Center for Statistics and Data Science.

Logistics

- Lectures: Tuesdays & Thursdays 1:00 -2:30am
- **Optional Recitation**: TBD.
- Homework: weekly. Total 11, 10 best kept (30%).
- Midterm: Nov. 8, in class, 1 hours and 20 minutes (30 %). Closed books closed notes. Cheatsheet.
- Final: TBD, 2 hours (40%). Open books, open notes.

Miscellaneous

- Prerequisites: Probability (18.600 or 6.041), Calculus 2, notions of linear algebra (matrix, vector, multiplication, orthogonality,...)
- Reading: There is no required textbook
- Slides are posted on course website

https://ocw.mit.edu/courses/mathematics/18-650-statistics-for-applications-fall-2016/lecture-slides

- **Videolectures:** Each lecture is recorded and posted online. Attendance is still recommended.

Why statistics?

Not only in the press

Hydrology Netherlands, 10th century, building dams and dykes Should be high enough for most floods Should not be too expensive (high)

Insurance Given your driving record, car information, coverage. What is a fair premium?

Clinical trials A drug is tested on 100 patients; 56 were cured and 44 showed no improvement. Is the drug effective?

Randomness

What is common to all these examples?

Randomness

What is common to all these examples?

RANDOMNESS

Randomness

What is common to all these examples?

RANDOMNESS

Associated questions:

- Notion of average ("**fair** premium", ...)
- Quantifying chance ("**most of** the floods", ...)
- Significance, variability, ...

Probability

- Probability studies randomness (hence the prerequisite)
- Sometimes, the physical process is completely known: dice, cards, roulette, fair coins, ...

Examples

Rolling 1 die:

- Alice gets \$1 if # of dots ≤ 3
- Bob gets \$2 if # of dots ≤ 2

Who do you want to be: Alice or Bob?

Rolling 2 dice:

- Choose a number between 2 and 12
- Win \$100 if you chose the sum of the 2 dice

Which number do you choose?
Well known random process from physics: 1/6 chance of each side, dice are independent. We can deduce the probability of outcomes, and expected \$ amounts. This is **probability**.

Probability

- Probability studies randomness (hence the prerequisite)
- Sometimes, the physical process is completely known: dice, cards, roulette, fair coins, ...

Examples

Rolling 1 die:

- Alice gets $1 if # of dots 3
- Bob gets $2 if # of dots 2

Who do you want to be: Alice or Bob?

Rolling 2 dice:

- Choose a number between 2 and 12
- Win $100 if you chose the sum of the 2 dice

Which number do you choose?

Well known random process from physics: 1/6 chance of each side, dice are independent. We can deduce the probability of outcomes, and expected $ amounts. This is **probability**.

Probability

- Probability studies randomness (hence the prerequisite)
- Sometimes, the physical process is completely known: dice, cards, roulette, fair coins, ...

Examples

Rolling 1 die:

- Alice gets $1 if # of dots 3
- Bob gets $2 if # of dots 2

Who do you want to be: Alice or Bob?

Rolling 2 dice:

- Choose a number between 2 and 12
- Win $100 if you chose the sum of the 2 dice

Which number do you choose?

Well known random process from physics: 1/6 chance of each side, dice are independent. We can deduce the probability of outcomes, and expected $ amounts. This is **probability**.

Probability

- Probability studies randomness (hence the prerequisite)
- Sometimes, the physical process is completely known: dice, cards, roulette, fair coins, ...

Examples

Rolling 1 die:

- Alice gets \$1 if # of dots 3
- Bob gets \$2 if # of dots 2

Who do you want to be: Alice or Bob?

Rolling 2 dice:

- Choose a number between 2 and 12
- Win \$100 if you chose the sum of the 2 dice

Which number do you choose?

Well known random process from physics: $1/6$ chance of each side, dice are independent. We can deduce the probability of outcomes, and expected \$ amounts. This is **probability**.

Statistics and modeling

- How about more complicated processes? Need to estimate parameters from data. This is **statistics**
- Sometimes real randomness (random student, biased coin, measurement error, ...)
- Sometimes deterministic but too complex phenomenon: **statistical modeling**

  $$\text{Complicated process "=" Simple process + random noise}$$

- (good) Modeling consists in choosing (plausible) simple process **and** noise distribution.

Statistics vs. probability

Probability Previous studies showed that the drug was 80% effective. Then we can anticipate that for a study on 100 patients, in average 80 will be cured and at least 65 will be cured with 99.99% chances.

Statistics Observe that 78/100 patients were cured. We (will be able to) conclude that we are 95% confident that for other studies the drug will be effective on between 69.88% and 86.11% of patients

18.650

What this course is about

- Understand **mathematics** behind statistical methods
- Justify quantitive statements given modeling assumptions
- Describe interesting mathematics arising in statistics
- Provide a math toolbox to extend to other models.

What this course is **not** about

- Statistical thinking/modeling (applied stats, e.g. IDS.012)
- Implementation (computational stats, e.g. IDS.012)
- Laundry list of methods (boring stats, e.g. AP stats)

18.650

What this course is about

- Understand **mathematics** behind statistical methods
- Justify quantitive statements given modeling assumptions
- Describe interesting mathematics arising in statistics
- Provide a math toolbox to extend to other models.

What this course is **not** about

- Statistical thinking/modeling (applied stats, e.g. IDS.012)
- Implementation (computational stats, e.g. IDS.012)
- Laundry list of methods (boring stats, e.g. AP stats)

Let's do some statistics

Heuristics (1)

"A neonatal right-side preference makes a surprising romantic reappearance later in life."

- Let p denote the proportion of couples that turn their head to the right when kissing.
- Let us design a statistical experiment and analyze its outcome.
- Observe n kissing couples times and collect the value of each outcome (say 1 for RIGHT and 0 for LEFT);
- *Estimate* p with the proportion $\hat{p}$ of RIGHT.
- Study: "Human behaviour: Adult persistence of head-turning asymmetry" (Nature, 2003): $n = 124$, 80 to the right so

$$\hat{p} = \frac{80}{124} = 64.5\%$$

Heuristics (2)

Back to the data:

- 64.5% is much larger than 50% so there seems to be a preference for turning right.
- What if our data was RIGHT, RIGHT, LEFT ($n = 3$). That's 66.7% to the right. Even better?
- Intuitively, we need a large enough sample size n to make a call. How large?

We need **mathematical modeling** to understand the accuracy of this procedure?

Heuristics (3)

Formally, this procedure consists of doing the following:

- For $i = 1, \ldots, n$, define $R_i = 1$ if the ith couple turns to the right RIGHT, $R_i = 0$ otherwise.
- The estimator of p is the sample average

$$\hat{p} = \bar{R}_n = \frac{1}{n} \sum_{i=1}^{n} R_i.$$

What is the accuracy of this estimator ?

In order to answer this question, we propose a statistical model that describes/approximates well the experiment.

Heuristics (4)

Coming up with a model consists of making assumptions on the observations $R_i, i = 1, \ldots, n$ in order to draw statistical conclusions. Here are the assumptions we make:

1. Each R_i is a random variable.
2. Each of the r.v. R_i is Bernoulli with parameter p.
3. $R_1, \ldots, R_n$ are mutually independent.

Heuristics (5)

Let us discuss these assumptions.

1. Randomness is a way of modeling lack of information; with perfect information about the conditions of kissing (including what goes in the kissers' mind), physics or sociology would allow us to predict the outcome.

2. Hence, the R_i's are necessarily Bernoulli r.v. since $R_i \in \{0, 1\}$. They could still have a different parameter $R_i \sim \text{Ber}(p_i)$ for each couple but we don't have enough information with the data estimate the p_i's accurately. So we simply assume that our observations come from the same process: $p_i = p$ for all i

3. Independence is reasonable (people were observed at different locations and different times).

Two important tools: LLN & CLT

Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. r.v., $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{V}[X]$.

- Laws of large numbers (weak and strong):

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P},\text{ a.s.}} \mu.$$

- Central limit theorem:

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

(Equivalently, $\sqrt{n}\,(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,\sigma^2)$.)

Consequences (1)

- The LLN's tell us that

$$\bar{R}_n \xrightarrow[n\to\infty]{\mathbb{P},\ \text{a.s.}} p.$$

- Hence, when the size n of the experiment becomes large, $\bar{R}_n$ is a *good* (say "*consistent*") estimator of p.

- The CLT refines this by quantifying *how good* this estimate is.

Consequences (2)

$\Phi(x)$: cdf of $\mathcal{N}(0,1)$;

$\Phi_n(x)$: cdf of $\sqrt{n}\,\dfrac{\bar{R}_n - p}{\sqrt{p(1-p)}}$.

CLT: $\Phi_n(x) \approx \Phi(x)$ when n becomes large. Hence, for all $x > 0$,

$$\mathbb{P}\left[|\bar{R}_n - p| \geq x\right] \approx 2\left(1 - \Phi\left(\frac{x\sqrt{n}}{\sqrt{p(1-p)}}\right)\right).$$

Consequences (3)

Consequences:

- Approximation on how $\bar{R}_n$ concentrates around p;
- For a fixed $\alpha \in (0,1)$, if $q_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$, then with probability $\approx 1-\alpha$ (if n is large enough !),

$$\bar{R}_n \in \left[p - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, p + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}\right].$$

Consequences (4)

- Note that no matter the (unknown) value of p,

$$p(1-p) \quad 1/4.$$

- Hence, roughly with probability at least $1-\alpha$,

$$\bar{R}_n \in \left[p - \frac{q_{\alpha/2}}{2\sqrt{n}}, p + \frac{q_{\alpha/2}}{2\sqrt{n}}\right].$$

- In other words, when n becomes large, the interval $\left[\bar{R}_n - \frac{q_{\alpha/2}}{2\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2}}{2\sqrt{n}}\right]$ contains p with probability $\geq 1-\alpha$.
- This interval is called an *asymptotic confidence interval* for p.
- In the kiss example, we get

$$\left[0.645 \pm \frac{1.96}{2\sqrt{124}}\right] = [0.56, 0.73]$$

If the extreme ($n = 3$ case) we would have $[0.10, 1.23]$ but CLT is not valid! Actually we can make exact computations!

Another useful tool: Hoeffding's inequality

What if n is not so large ?

Hoeffding's inequality (i.i.d. case)

Let n be a positive integer and $X, X_1, \ldots, X_n$ be i.i.d. r.v. such that $X \in [a, b]$ a.s. ($a < b$ are given numbers). Let $\mu = \mathbb{E}[X]$. Then, for all $\varepsilon > 0$,

$$\mathbb{P}[|\bar{X}_n - \mu| \geq \varepsilon] \quad 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}.$$

Consequence:

- For $\alpha \in (0, 1)$, with probability $\geq 1 - \alpha$,

$$\bar{R}_n - \sqrt{\frac{\log(2/\alpha)}{2n}} \quad p \quad \bar{R}_n + \sqrt{\frac{\log(2/\alpha)}{2n}}.$$

- This holds even for small sample sizes n.

Review of different types of convergence (1)

Let $(T_n)_{n\geq 1}$ a sequence of r.v. and T a r.v. (T may be deterministic).

- Almost surely (a.s.) convergence:

$$T_n \xrightarrow[n\to\infty]{\text{a.s.}} T \quad \text{iff} \quad \mathbb{P}\left[\left\{\omega : T_n(\omega) \xrightarrow[n\to\infty]{} T(\omega)\right\}\right] = 1.$$

- Convergence in probability:

$$T_n \xrightarrow[n\to\infty]{\mathbb{P}} T \quad \text{iff} \quad \mathbb{P}\left[|T_n - T| \geq \varepsilon\right] \xrightarrow[n\to\infty]{} 0, \quad \forall \varepsilon > 0.$$

Review of different types of convergence (2)

- Convergence in L^p $(p \geq 1)$:

$$T_n \xrightarrow[n\to\infty]{L^p} T \quad \text{iff} \quad \mathbb{E}\left[|T_n - T|^p\right] \xrightarrow[n\to\infty]{} 0.$$

- Convergence in distribution:

$$T_n \xrightarrow[n\to\infty]{(d)} T \quad \text{iff} \quad \mathbb{P}[T_n \quad x] \xrightarrow[n\to\infty]{} \mathbb{P}[T \quad x],$$

for all $x \in \mathbb{R}$ at which the cdf of T is continuous.

Remark

These definitions extend to random vectors (i.e., random variables in $\mathbb{R}^d$ for some $d \geq 2$).

Review of different types of convergence (3)

Important characterizations of convergence in distribution

The following propositions are equivalent:

(i) $T_n \xrightarrow[n\to\infty]{(d)} T$;

(ii) $\mathbb{E}[f(T_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[f(T)]$, for all continuous and bounded function f;

(iii) $\mathbb{E}\left[e^{ixT_n}\right] \xrightarrow[n\to\infty]{} \mathbb{E}\left[e^{ixT}\right]$, for all $x \in \mathbb{R}$.

Review of different types of convergence (4)

Important properties

- If $(T_n)_{n\geq 1}$ converges a.s., then it also converges in probability, and the two limits are equal a.s.

- If $(T_n)_{n\geq 1}$ converges in L^p, then it also converges in L^q for all q p and in probability, and the limits are equal a.s.

- If $(T_n)_{n\geq 1}$ converges in probability, then it also converges in distribution

- If f is a continuous function:

$$T_n \xrightarrow[n\to\infty]{\text{a.s.}/\mathbb{P}/(d)} T \quad \Rightarrow \quad f(T_n) \xrightarrow[n\to\infty]{\text{a.s.}/\mathbb{P}/(d)} f(T).$$

Review of different types of convergence (6)

Limits and operations

One can add, multiply, ... limits almost surely and in probability. If $U_n \xrightarrow[n\to\infty]{\text{a.s.}/\mathbb{P}} U$ and $V_n \xrightarrow[n\to\infty]{\text{a.s.}/\mathbb{P}} V$, then:

- $U_n + V_n \xrightarrow[n\to\infty]{\text{a.s.}/\mathbb{P}} U + V$,
- $U_n V_n \xrightarrow[n\to\infty]{\text{a.s.}/\mathbb{P}} UV$,
- If in addition, $V \neq 0$ a.s., then $\frac{U_n}{V_n} \xrightarrow[n\to\infty]{\text{a.s.}/\mathbb{P}} \frac{U}{V}$.

⚠ In general, these rules **do not** apply to convergence in distribution unless the **pair** (U_n, V_n) converges in distribution to (U, V).

Another example (1)

- You observe the times between arrivals of the T at Kendall: $T_1, \ldots, T_n$.
- You **assume** that these times are:
  - Mutually independent
  - Exponential random variables with common parameter $\lambda > 0$.
- You want to *estimate* the value of λ, based on the observed arrival times.

Another example (2)

Discussion of the assumptions:

- Mutual independence of $T_1, \dots, T_n$: plausible but not completely justified (often the case with independence).
- $T_1, \dots, T_n$ are exponential r.v.: **lack of memory** of the exponential distribution:

  $$\mathbb{P}[T_1 > t + s | T_1 > t] = \mathbb{P}[T_1 > s], \quad \forall s, t \geq 0.$$

  Also, $T_i > 0$ almost surely!
- The exponential distributions of $T_1, \dots, T_n$ have the same parameter: in average all the same inter-arrival time. True only for limited period (rush hour $\neq$ 11pm).

Another example (3)

- Density of T_1:

$$f(t) = \lambda e^{-\lambda t}, \quad \forall t \geq 0.$$

- $\mathbb{E}[T_1] = \frac{1}{\lambda}$.

- Hence, a natural estimate of $\frac{1}{\lambda}$ is

$$\bar{T}_n := \frac{1}{n} \sum_{i=1}^{n} T_i.$$

- A natural estimator of λ is

$$\hat{\lambda} := \frac{1}{\bar{T}_n}.$$

Another example (4)

- By the LLN's,

$$\bar{T}_n \xrightarrow[n\to\infty]{\text{a.s.}/\mathbb{P}} \frac{1}{\lambda}$$

- Hence,

$$\hat{\lambda} \xrightarrow[n\to\infty]{\text{a.s.}/\mathbb{P}} \lambda.$$

- By the CLT,

$$\sqrt{n}\,\left(\bar{T}_n - \frac{1}{\lambda}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \lambda^{-2}).$$

- How does the CLT transfer to $\hat{\lambda}$? How to find an asymptotic confidence interval for λ ?

The Delta method

Let $(Z_n)_{n\geq 1}$ sequence of r.v. that satisfies

$$\sqrt{n}(Z_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2),$$

for some $\theta \in \mathbb{R}$ and $\sigma^2 > 0$ (the sequence $(Z_n)_{n\geq 1}$ is said to be *asymptotically normal around* θ).

Let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable at the point θ. Then,

- $(g(Z_n))_{n\geq 1}$ is also asymptotically normal;
- More precisely,

$$\sqrt{n}\,(g(Z_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, g'(\theta)^2\sigma^2).$$

Consequence of the Delta method (1)

- $\sqrt{n}\left(\hat{\lambda}-\lambda\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,\lambda^2)$.

- Hence, for $\alpha \in (0,1)$ and when n is large enough,

$$|\hat{\lambda}-\lambda| \quad \frac{q_{\alpha/2}\lambda}{\sqrt{n}}.$$

- Can $\left[\hat{\lambda}-\frac{q_{\alpha/2}\lambda}{\sqrt{n}}, \hat{\lambda}+\frac{q_{\alpha/2}\lambda}{\sqrt{n}}\right]$ be used as an asymptotic confidence interval for λ ?

- **No !** It depends on λ...

Consequence of the Delta method (2)

Two ways to overcome this issue:

- In this case, we can solve for λ:

$$|\hat{\lambda} - \lambda| \quad \frac{q_{\alpha/2}\lambda}{\sqrt{n}} \iff \lambda\left(1 - \frac{q_{\alpha/2}}{\sqrt{n}}\right) \quad \hat{\lambda} \quad \lambda\left(1 + \frac{q_{\alpha/2}}{\sqrt{n}}\right)$$
$$\iff \hat{\lambda}\left(1 + \frac{q_{\alpha/2}}{\sqrt{n}}\right)^{-1} \quad \lambda \quad \hat{\lambda}\left(1 - \frac{q_{\alpha/2}}{\sqrt{n}}\right)^{-1}.$$

  Hence, $\left[\hat{\lambda}\left(1 + \frac{q_{\alpha/2}}{\sqrt{n}}\right)^{-1}, \hat{\lambda}\left(1 - \frac{q_{\alpha/2}}{\sqrt{n}}\right)^{-1}\right]$ is an asymptotic confidence interval for λ.

- A systematic way: *Slutsky's theorem*.

Slutsky's theorem

Slutsky's theorem

Let $(X_n), (Y_n)$ be two sequences of r.v., such that:

(i) $X_n \xrightarrow[n\to\infty]{(d)} X$;

(ii) $Y_n \xrightarrow[n\to\infty]{\mathbb{P}} c$,

where X is a r.v. and c is a given real number. Then,

$$(X_n, Y_n) \xrightarrow[n\to\infty]{(d)} (X, c).$$

In particular,

$$X_n + Y_n \xrightarrow[n\to\infty]{(d)} X + c,$$

$$X_n Y_n \xrightarrow[n\to\infty]{(d)} cX,$$

$$\dots$$

Consequence of Slutsky's theorem (1)

- Thanks to the Delta method, we know that

$$\sqrt{n}\,\frac{\hat{\lambda}-\lambda}{\lambda} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

- By the weak LLN,

$$\hat{\lambda} \xrightarrow[n\to\infty]{\mathbb{P}} \lambda.$$

- Hence, by Slutsky's theorem,

$$\sqrt{n}\,\frac{\hat{\lambda}-\lambda}{\hat{\lambda}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

- Another asymptotic confidence interval for λ is

$$\left[\hat{\lambda}-\frac{q_{\alpha/2}\hat{\lambda}}{\sqrt{n}},\hat{\lambda}+\frac{q_{\alpha/2}\hat{\lambda}}{\sqrt{n}}\right].$$

Consequence of Slutsky's theorem (2)

Remark:

- In the first example (kisses), we used a problem dependent trick: “$p(1-p) \quad 1/4$”.

- We could have used Slutsky's theorem and get the asymptotic confidence interval

$$\left[\bar{R}_n - \frac{q_{\alpha/2}\sqrt{\bar{R}_n(1-\bar{R}_n)}}{\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2}\sqrt{\bar{R}_n(1-\bar{R}_n)}}{\sqrt{n}}\right].$$