Statistics for Applications

Chapter 1: Introduction

Introduction

- My webpage: `http://math.mit.edu/~vebrunel/`
- Aims of this course:
  - 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.
- No required textbook.
- Office hours: Wednesdays, 4-6pm, office 2-239b.

Work required from the students

- Six graded problem sets (20% of the final grade): theoretical exercises and programming (in R language).
- Weekly non graded (but highly recommended) exercises.
- In-class midterm exam on Thursday March 17 (30% of the final grade): theoretical problems.
- Final exam (50% of the final grade): 2 hours, location and time TBD.

Let's get started with an introduction to statistics.

Heuristics (1)

- You want to measure the parameter p associated to a coin that is in your possession;
- Let us design a statistical experiment and analyze its outcome.
- You toss the coin many (say, n) times and collect the value of each outcome;
- You *estimate* p with the proportion of Heads within all the outcomes.

What guarantees the validity of this procedure ?

Heuristics (2)

Formally, this procedure consists of doing the following:

- For $i = 1, \dots, n$, define $H_i = 1$ if Heads showed up at the i-th toss, $H_i = 0$ otherwise.
- The estimator of p is the sample average

$$\bar{H}_n = \frac{1}{n} \sum_{i=1}^{n} H_i.$$

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 (3)

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

1. Each H_i is a random variable.
2. Each of the r.v. H_i is Bernoulli with parameter p.
3. $H_1, \ldots, H_n$ are mutually independent.

Heuristics (4)

Let us discuss these assumptions.

1. Randomness is a way of modeling lack of information; with perfect information about the conditions of flipping the coin, physics would allow to predict all the outcomes.

2. Hence, the H_i's are necessarily Bernoulli r.v. since $H_i \in \{0, 1\}$. Their parameter would be p if the coin could not land on its side... See `https://www.seas.harvard.edu/softmat/downloads/2011-10.pdf` for a nice discussion.

3. Independence is reasonable if there is no change in the way of tossing the coin (e.g., no learning process).

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

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

- Hence, when the size n of the experiment becomes large, $\bar{H}_n$ is a *good* (say "*consistent*") estimate 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 $\sqrt{n}\,\frac{\bar{H}_n - p}{\sqrt{p(1-p)}}$.

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

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

Consequences (3)

Consequences:

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

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

Consequences (4)

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

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

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

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

- In other words, when n becomes large, the interval $\left[\bar{H}_n - \frac{q_\alpha}{2\sqrt{n}}, \bar{H}_n + \frac{q_\alpha}{2\sqrt{n}} \right]$ contains p with probability $\geq 1-\alpha$.

- This interval is called an *asymptotic confidence interval* for p.

- What if n is not so large ?

Another useful tool: Hoeffding's inequality

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 $\varepsilon > 0$,

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

Consequence:

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

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

- 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:

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

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 $q \leq p$ and in probability, and the limits are equal a.s.

- 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 (5)

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 $\dfrac{U_n}{V_n} \xrightarrow[n\to\infty]{\text{a.s.}/\mathbb{P}} \dfrac{U}{V}$.

⚠ 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 new individuals in a queue (e.g., at a call center): $T_1, \ldots, T_n$.
- You **assume** that these times are:
  - Mutually independent
  - Exponential random variables with some 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 $T_1, \ldots, T_n$: the individuals are not related to each other, hence, do not decide when to arrive based on others' arrival times.

- $T_1, \ldots, T_n$ are exponential r.v.: **lack of memory** of the exponential distribution.

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

- The exponential distributions of $T_1, \ldots, T_n$ have the same parameter: homogeneous behavior in the population.

Another example (3)

- Density of T_1:

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

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

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

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

- 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

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

for some $\vartheta \in \mathbb{R}$ and $\sigma^2 > 0$ (the sequence $(Z_n)_{n\geq 1}$ is called *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,

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

Consequence of the Delta method (1)

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

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

$$|\hat{\lambda}-\lambda| \le \frac{q_\alpha \lambda}{\sqrt{n}}.$$

- Can $\left[\hat{\lambda}-\frac{q_\alpha\lambda}{\sqrt{n}}, \hat{\lambda}+\frac{q_\alpha\lambda}{\sqrt{n}}\right]$ 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:

- A problem-dependent way:

$$|\hat{\lambda} - \lambda| \leq \frac{q_\alpha \lambda}{\sqrt{n}} \iff \lambda\left(1 - \frac{q_\alpha}{\sqrt{n}}\right) \leq \hat{\lambda} \leq \lambda\left(1 + \frac{q_\alpha}{\sqrt{n}}\right)$$

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

  Hence, $\left[\hat{\lambda}\left(1 + \frac{q_\alpha}{\sqrt{n}}\right)^{-1}, \hat{\lambda}\left(1 - \frac{q_\alpha}{\sqrt{n}}\right)^{-1}\right]$ 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,

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

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

$$\ldots$$

Consequence of Slutsky's theorem (1)

- Thanks to the Delta method, we know that

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

- By the weak LLN,

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

- Hence, by Slutsky's theorem,

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

- Another asymptotic confidence interval for λ is

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

Consequence of Slutsky's theorem (2)

Remark:

- In the first example (coin tosses), we used a problem dependent way: "$p(1-p) \leq 1/4$".
- We could have used Slutsky's theorem and get the asymptotic confidence interval

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