

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**.

* To understand randomness we have to go through Probability.

* Probability is the study of randomness

* Probability is what we are gonna use to describe the randomness. We will see how it interacts with Statistics.

* In Probability randomness is very well understood. When we saw a Probability Problem, here was the chance of this happening, that happening etc..

For example, what's the prob that I watch GOT or have YouTube Premium in that much. We have all these crazy questions, But you were able to build them, But all the basic numbers were already given to you. Statistics will be about finding those basic numbers

* In Probability, we know what's the Probability of Head's, or Prob of getting 1 in a dice

* Now if I was a statistician, statistician would be the guy just arrived on earth, had never seen a die and need's to understand that a die come with $\frac{1}{6}$ on each side and the way he would do it is just roll the die until he get some counts and try to estimate those thing's.

* Or maybe that guy come and tell actually
The $P(1) = \frac{1}{6} + 0.001$ & $P(2) = \frac{1}{6} - 0.05, \dots$
there would be some fluctuation's on these

* And his role as statistician is to say, Listen this is too complicated model for this thing and these should be the same number's for all Number's (That's part of modelling)

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

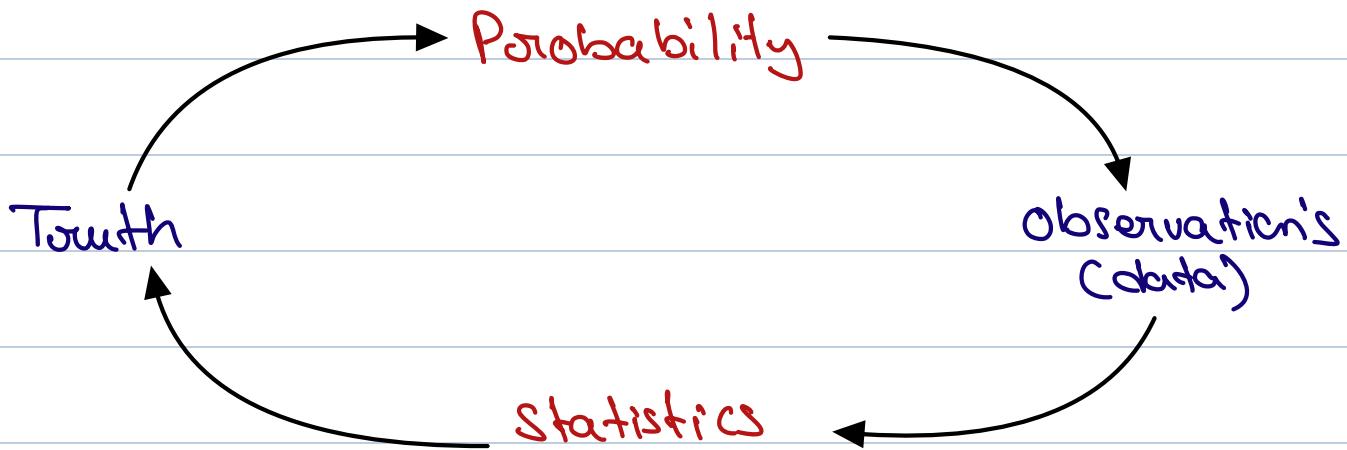
- ⌚ ▶ How about more complicated processes? Need to estimate parameters from data. This is **statistics**
- ▶ Sometimes real randomness (random student, biased coin, measurement error, ...)
- ▶ Sometimes deterministic but too complex phenomenon:
statistical modeling
 - Complicated process “=” Simple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

Statistics vs Probability

Circle of trust



When we do Probability, we are given the truth, somebody tells you what die and in Rolling. We know exactly parameters of the problem's are, and what we are trying to do is to describe what the outcome's are going to be. We can say, for fair die $P(i) = 1/6$.

So we know from the truth, from the distribution of the random variable that does the data gonna look like. Statistics is about going BACKwards

* if I have some data , what was the tooth that generated , And since there are so many possible teeth's , modelling say's we have to pick one of those simpler possible teeths so that we can average out. "Statistics basically mean's averaging."

18.650

What this course is about

- ▶ Understand **mathematics** behind statistical methods
- ▶ Justify quantitative 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 quantitative 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 $i = 1, \dots, n$, define $R_i = 1$ if the i th 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.

$$\text{The Estimator } \hat{P} = \bar{R}_n = \frac{1}{n} \sum_{i=1}^n R_i$$

This is an Estimator, so an Estimator is different from Estimate.

\Rightarrow Estimate is a number, here my Estimate is 64.5.

\Rightarrow My Estimator is $\bar{R}_n = \hat{P} = \frac{1}{n} \sum_{i=1}^n R_i$, where I keep all the variables free. In particular I keep those variable to be random. (random couple kissing left or right in the outcome of random process, Just like flipping a coin, getting heads or tails)

$\Rightarrow R_i$ is a random variable

$\Rightarrow \bar{R}$ is a average is a random variable

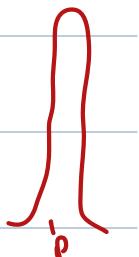
So An Estimator is a random variable and Estimate is the realization of the random variable or in other words the

Value we get for this random variable (\bar{R}) once we plug in the number's that are collected.

What is the accuracy of this Estimator?

What would we want from an estimator, may be we won't want it to fluctuate too much. As its a giv. maybe we don't want it to be too volatile.

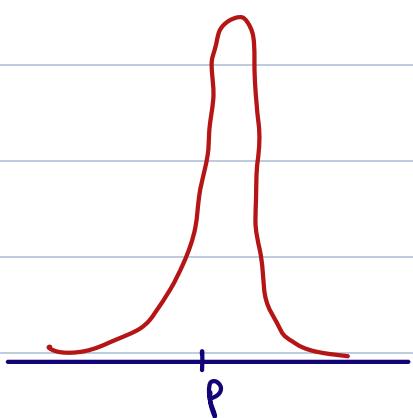
- * we could have one estimator, which just throw out 122 couples and keep only 2 couples and average those two numbers (That's definitely worst estimator, than keeping all the 120 observations)
- * so we need to find a way to say that (fluctuating) (we want a measure on how much it fluctuate)
- * we want density to be pretty narrow, and close to actual value.



Another Estimator: $\hat{P} = 72\%$ (constant)

that's an Estimator (constant r.v), tell's go out there just do your favorite study about drug performance and then just tell 72%. (It's a Stupid Estimator, But it's an Estimator). This Estimator is not volatile at all. Everytime we are gonna have new study, it still going to be 70%
 $\text{Var}(\bar{R}) = 0$ (Beautiful, BUT the problem is its not gonna be very close to the actual value)

\Rightarrow so we need an Estimator to be random variable, we want density to be pretty narrow (low variance), close to actual value (Bias)



Heuristics (4) 99.9% of statistics is all about
replacing expectation's with averages.

Coming up with a model consists of making assumptions on the observations $R_i, i = 1, \dots, n$ 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, \dots, 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, \dots, 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):

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

- ▶ Central limit theorem:

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

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

x_1, x_2, \dots, x_n iid $Ber(p)$

p unknown

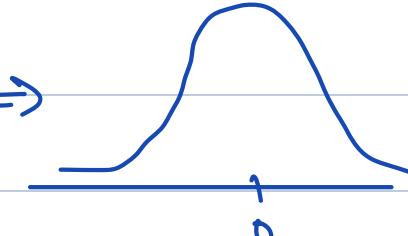
* P is the expectation of x_i , well if I need to estimate the proportion of people who turn their head to the right when they kiss, I just basically we are gonna compute the average.

$$\hat{p} = \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i,$$

Now we want to build confidence intervals around this. How much that this \hat{p} fluctuates. x_i is r.v, $\hat{p} = \bar{x}$ is a random variable (average of random variables), so we want to know what its distribution is. And if we know what's the distribution is then we actually know, where it fluctuates? what the expectation is? And which value it tends to fluctuate?

$$\sqrt{n} \left(\frac{\bar{x}_n - p}{\sqrt{P(\bar{x}-p)}} \right) \xrightarrow[n \rightarrow \infty]{(d)} N(0,1) \quad [\text{CLT}]$$

$\exists t \cdot V \Rightarrow \left[\begin{array}{l} \text{when } n \text{ is large} \\ \text{enough it looks like} \\ N(0,1) \end{array} \right] \Rightarrow$



when it perfectly looks like $N(0,1)$, so we can understand perfectly its fluctuation's.

$$P(N(0,1) > q_{\alpha/2}) = \alpha$$

|| because of CLT

$$P \left(\left| \sqrt{n} \left(\frac{\bar{x} - p}{\sqrt{P(\bar{x}-p)}} \right) \right| > q_{\alpha/2} \right) \approx \alpha$$

↳ solve for p

What was important in the building of this confidence interval?

When we take this function. $\left[\frac{\bar{X} - P}{\sqrt{P(1-P)}} \right] \text{ in}$

The CLT tell's us that it converges to $N(0,1)$ (Something that we know)

Most Important: is that it does not depend on anything that I don't know.

Consequences (1)

- ▶ The LLN's tell us that

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

- ▶ 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 $\sqrt{n} \frac{\bar{R}_n - p}{\sqrt{p(1-p)}}$.

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

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

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 !),

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

Assume I have 3 r.v's

X_1, X_2, X_3 iid Bernoulli(P)

P = unknown's

then the possible outcomes are

$$0 \ 0 \ 0 \rightarrow 0$$

$$\begin{matrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{matrix} \rightarrow \frac{1}{3}$$

$$\begin{matrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{matrix} \rightarrow \frac{2}{3}$$

$$1 \ 1 \ 1 \rightarrow 1$$

Possible values for $\bar{X}_3 = \left\{ 0, \frac{1}{3}, \frac{2}{3}, 1 \right\}$

then what's the probability of

$IP(\bar{X}_3 = 0), IP(\bar{X}_3 = \frac{1}{3}), IP(\bar{X}_3 = \frac{2}{3}), IP(\bar{X}_3 = 1)$?

These values depend on the unknown P value.

if $P=0$, then $IP(\bar{X}_3=0)=1$, $IP(\bar{X}_3=\frac{1}{3})=0$
 $IP(\bar{X}_3=\frac{2}{3})=0$
 $IP(\bar{X}_3=1)=0$

if $P=\frac{1}{2}$, then

$$IP(\bar{X}_3=0)=\frac{1}{8}, \quad IP(\bar{X}_3=\frac{2}{3})=IP(\bar{X}_3=\frac{1}{3})=\frac{3}{8}$$

$$IP(\bar{X}_3=1)=\frac{1}{8}$$

these value depend on P (change with
 value of P). But we don't know
 P value.

$$IP(\bar{X}_3=0)=(1-P)^3, \quad IP(\bar{X}_3=\frac{1}{3})=(1-P)^2P$$

$$IP(\bar{X}_3=\frac{2}{3})=P^2(1-P), \quad IP(\bar{X}_3=1)=P^3$$

$$\bar{X}_3 \sim \text{Binomial}(3, P)$$

Consequences (4)

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

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

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

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

- ▶ 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, \dots, 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 $\varepsilon > 0$,

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

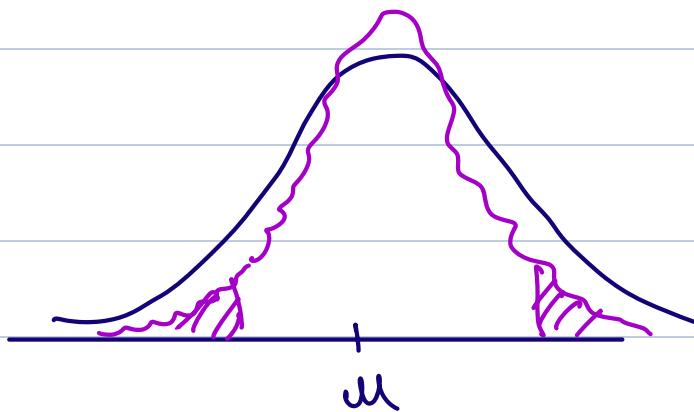
Consequence:

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

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

- ▶ This holds even for small sample sizes n .

Essentially what Hoeffding's inequality telling us that,



Assume we have a mean μ . What Hoeffding's inequality is telling me is that, if we take the average of bounded r.v, then their PDF or PMF is gonna be the tails are bounded by Gaussian.

$$\Rightarrow \underbrace{\Pr(|\bar{X}_n - \mu| > \varepsilon)}_{\text{tail probability}} \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}$$

is bounded by something similar to Gaussian.

$$e^{-\frac{2n\varepsilon^2}{(b-a)^2}} = e^{-\frac{2\varepsilon^2}{(\frac{b-a}{\sqrt{n}})^2}}$$

what Hoeffding's tell us bermoulli $[0,1]$
 (a,b)

$$IP(|\bar{X}_n - P| > \varepsilon) \leq 2e^{-2n\varepsilon^2} = \alpha$$

$$2e^{-2n\varepsilon^2} = \alpha \Rightarrow \varepsilon = \sqrt{\frac{\log(\frac{2}{\alpha})}{2n}}$$

$$\Rightarrow IP\left(\bar{X}_n - \sqrt{\frac{\log(\frac{2}{\alpha})}{2n}} \leq P \leq \bar{X}_n + \sqrt{\frac{\log(\frac{2}{\alpha})}{2n}}\right) \leq \alpha$$

$$|\bar{X}_n - P| < \varepsilon$$

$$\Rightarrow -\varepsilon \leq P - \bar{X}_n < \varepsilon$$

$$\Rightarrow \bar{X}_n - \varepsilon < P < \bar{X}_n + \varepsilon$$

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:

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

- ▶ Convergence in probability:

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

Review of different types of convergence (2)

- ▶ Convergence in L^p ($p \geq 1$):

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

- ▶ Convergence in distribution:

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

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:

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

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

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

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:

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

Review of different types of convergence (6)

Limits and operations

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

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



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, \dots, 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,

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

- ▶ Hence,

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

- ▶ By the CLT,

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

- ▶ 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

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

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 $g : \mathbb{R} \rightarrow \mathbb{R}$ be continuously differentiable at the point θ .

Then,

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

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

Consequence of the Delta method (1)

- ▶ $\sqrt{n} \left(\hat{\lambda} - \lambda \right) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \lambda^2).$
- ▶ Hence, for $\alpha \in (0, 1)$ and when n is large enough,

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

- ▶ 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 :

$$\begin{aligned} |\hat{\lambda} - \lambda| &= \frac{q_{\alpha/2}\lambda}{\sqrt{n}} \iff \lambda \left(1 - \frac{q_{\alpha/2}}{\sqrt{n}}\right) \leq \hat{\lambda} \leq \lambda \left(1 + \frac{q_{\alpha/2}}{\sqrt{n}}\right) \\ &\iff \hat{\lambda} \left(1 + \frac{q_{\alpha/2}}{\sqrt{n}}\right)^{-1} \leq \lambda \leq \hat{\lambda} \left(1 - \frac{q_{\alpha/2}}{\sqrt{n}}\right)^{-1}. \end{aligned}$$

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:

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

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

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

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

In particular,

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

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

...

Consequence of Slutsky's theorem (1)

- ▶ Thanks to the Delta method, we know that

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

- ▶ By the weak LLN,

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

- ▶ Hence, by Slutsky's theorem,

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

- ▶ 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: “ $p(1 - p) = 1/4$ ”.
- ▶ We could have used Slutsky's theorem and get the asymptotic confidence interval

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

MIT OpenCourseWare
<https://ocw.mit.edu>

18.650 / 18.6501 Statistics for Applications

Fall 2016

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.