GM – Probabilité et Statistique

http://moodle.epfl.ch/course/view.php?id=18431

Lecture 6

- Jointly distributed RVs
- Independent RVs
- Sums of independent RVs
- Sampling
- Sampling distribution of a statistic T
- Central Limit Theorem (CLT) / Théorème Central Limite (TCL)

Review : RVs (univariate)

- **Discrete RV** :
  1. probability mass function : $p(x) = P(X = x)$
  2. cumulative distribution function : $F(x) = P(X \leq x) = \sum_{i \leq x} p(i)$
- **Continuous RV : probability density function :**
  - $P(X \in B) = \int_B f(x)dx$
  - $f(x) \geq 0$ for each x
  - $\int_x f(x)\, dx = 1$
- **Continuous RV : cumulative distribution function :**
  - $F(x) = P(X \leq x) = \int_0^x f(u)\, du$
  - $F(-\infty) = 0$
  - $F(\infty) = 1$

Joint cumulative distribution function

- Until now we have considered only distributions of RVs *one at a time*
- In practice, it is often necessary to consider events concerning two (or even more) variables *simultaneously*
- To handle this type of problem, we define a Pour traiter de tels problèmes on définit une **joint cumulative distribution function F** for any pair of RVs X and Y :

$$F(a, b) = P(X \leq a, Y \leq b) \qquad -\infty < a, b < \infty$$

- Just like before, if we know the cumulative distribution function of a set of RVs (or the pmf or density), *we can address questions concerning probabilities*

Marginal cumulative distribution function

- The **marginal cumulative distribution function** for a RV is the cumulative distribution function of the single RV, *without regard* to the other RVs
- The cumulative distribution function of X is obtai-
ned from the joint cumulative distribution function of X and Y :

$$\begin{aligned} F_X(a) &= P(X \leq a) && \text{[definition]} \\ &= P(X \leq a, Y < \infty) && \text{[joint cdf]} \\ &= P(\lim_{b->\infty} X \leq a, Y \leq b) && \text{[subst. limit]} \\ &= \lim_{b->\infty} P(X \leq a, Y \leq b) && \text{[change order lim / P]} \\ &= \lim_{b->\infty} F(a, b) = F(a, \infty) && \text{[definition]} \end{aligned}$$

- Similarly, we find the cumulative distribution function of Y, $F_Y(b) = F(\infty, b)$

Joint probability mass function

- For *two discrete RVs* X and Y, we can define the **joint probability mass function** (joint pmf) as :

$$p(x,y) = P(X = x, Y = y)$$

- The **marginal pmf** of X can be obtained from the joint pmf $p(x,y)$:

$$\begin{aligned} p_X(x) &= P(X = x) \\ &= \sum_{y:p(x,y)>0} p(x,y), \end{aligned}$$

- *i.e.*, the marginal pmf of X is obtained by summing the joint pmf *over all possibilities of Y*
- The **marginal pmf** of Y is obtained similarly : $p_Y(y) = \sum_{x:p(x,y)>0} p(x,y)$

Joint probability density function

- The RVs X and Y are **jointly continuous** if there is a function $f(x, y)$ defined for all real x and y such that for every set C of pairs of real numbers

$$P((X, Y) \in C) = \int \int_{(x,y)\in C} f(x, y)\, dx\, dy$$

- The function $f(x, y)$ is called the **joint probability density function** of X and Y
- Let A and B denote two sets of real numbers, $C = \{(x, y) : x \in A,\ y \in B)\}$; we have

$$P(X \in A, Y \in B) = \int_B \int_A f(x, y)\, dx\, dy$$

- The joint density function can be obtained from the joint cumulative distribution function by differentiation :

$$f(a, b) = \frac{\partial^2}{\partial a\, \partial b} F(a, b)$$

(where the partial derivatives are defined)

Marginal density

- For X and Y jointly continuous RVs, they are also *individually continuous*
- We obtain the **marginal density** of each RV as follows :

$$\begin{aligned} P(X \in A) &= P(X \in A,\ Y \in (-\infty, \infty)) \\ &= \int_A \left[\int_{-\infty}^{\infty} f(x, y)\, dy \right] dx = \int_A f_X(x)\, dx, \end{aligned}$$

where $f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$ is the (marginal) density of X

- Similarly, we obtain the (marginal) density of Y :

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

Example

Example 6.1 The joint density of X and Y is given by $f(x, y) = 2e^{-x}e^{-2y}$, $0 < x < \infty$, $0 < y < \infty$ ($f(x, y) = 0$ otherwise).

(a) $P(X > 1, Y < 1) =$

(b) $P(X < Y) =$

(c) $P(X < a) =$

Independent random variables

- We have already seen the concept of independence of *events*
- Now we define independence for *random variables*
- RVs X and Y are **independent** if for any two sets of real numbers A and B,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

- That is, the RVs X and Y are independent if the events $X \in A$ and $Y \in B$ are independent for all A and B

Independent random variables

- **Theorem :** RVs X and Y are independent *if and only if* the joint pmf (discrete RVs) or the joint density (continuous RVs) can be factored :

$$p_{X,Y}(x,y) = g(x)\,h(y) \qquad \text{for all } x \text{ and all } y;$$

$$f_{X,Y}(x,y) = g(x)\,h(y), \qquad -\infty < x < \infty,\ -\infty < y < \infty$$

- More generally, RVs $X_1, X_2, \ldots, X_n$ are **independent** if for any choice of n sets of real numbers $A_1, A_2, \ldots, A_n$,

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i)$$

Example

Example 6.2

(a) The joint density of X and Y is given by $f(x,y) = 6e^{-2x}e^{-3y},\ 0 < x < \infty,\ 0 < y < \infty\ (f(x,y) = 0$ otherwise). Are X and Y independent ??

(b) The joint density of X and Y is given by $f(x,y) = 24xy,\ 0 < x < 1,\ 0 < y < 1,\ 0 < x + y < 1\ (f(x,y) = 0$ otherwise). Are X and Y independent ??

Example

Example 6.3 If X and Y are independent Poisson RVs, $X \sim Pois(\lambda_1)$, $Y \sim Pois(\lambda_2)$, find the distribution of $X + Y$.

Solution The event $X + Y = n$ is the union of disjoint events $(X = k, Y = n - k)$ for $k = 0, 1, \ldots, n$; thus

$$
\begin{aligned}
P(X + Y = n) &= \sum_{k=0}^{n} P(X = k, Y = n - k) \\
&= \sum_{k=0}^{n} P(X = k) P(Y = n - k) \\
&= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!}
\end{aligned}
$$

Solution, cont.

$$
\begin{aligned}
&= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1 + \lambda_2)^n
\end{aligned}
$$

What distribution is this??

PAUSE

What is your IQ ? ?

IQ test

1. Does Father Christmas exist ?
2. Who is the best footballer in the world ?
3. Evaluate :

$$\int_{x=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Sampling

- The goal of a statistical study is to obtain some knowledge about a *population*, that is, **estimation of a parameter**
- Since a complete enumeration of the population is rarely practical, we need other, more practical means
- ⇒ **Sampling** consists of choosing among the population a certain number of individuals ('sampling units') for which we will obtain observations (data)
- Our data can be considered as arising from a *random process* : if we were to repeat the collection of the data, the results would be different, and this could influence any conclusions drawn based on the data
- That is, our conclusions are subject to *random variation*

Sampling distribution

- A **statistic** is a *function of the data*
- The (exact) distribution of a statistic T is called the **sampling distribution**
- The sampling distribution of a statistic is determined by the *sampling program (method)* – that is what defines the probability associated with each possible sample

Distribution of the sum of independent Normal RVs

- For VAs $X_1, \ldots, X_n$:

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n]$$

- For RVs $X_1, \ldots, X_n$ **independent** :

$$Var[X_1 + \cdots + X_n] = Var[X_1] + \cdots + Var[X_n]$$

- **Theorem** : Let $X_1, \ldots, X_n$ be **independent Normal** RVs with parameters $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$
- Then,

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Central Limit Theorem (CLT / TCL)

- The **Central Limit Theorem** is one of the most important results in probability/statistics, and is widely used as a problem-solving tool.
- **Theorem (CLT / TCL)** : Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) RVs, each having mean μ and variance σ^2. Then the distribution of

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the **standard normal distribution** as $n \to \infty$.
- In words, *the distribution of the sum (or average)* of a (sufficiently large) number of independent RVs is *approximately normal*.

Example

Example 6.4 An (imaginary) elevator has a maximum weight capacity of 3.6 metric tonnes (3600 kg). A certain population has an average weight of 70 kg, with an SD of 16 kg.

(a) What is the chance that a random sample of 49 (! !) people from this population overloads the elevator *? ?*

(b) Find the maximum number of people the elevator should accommodate in order that the chance of being overloaded is less that 1% **...**

Interval estimation

- Usually it is not very informative to give only a *point estimate* – a single number guess for the parameter value
- It is also of interest to have some idea of *the probable size of the error*
  - **standard error** (SE) : estimated SD of a parameter estimate
  - For example, $SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$, estimated by $SE = \frac{s}{\sqrt{n}}$
- Another way to present your estimate is in the form of a **confidence interval** (CI)

Deriving a CI for the population mean

the details are not part of the exam

- CLT : the *sampling distribution of the sample mean* is approximately Normal, with mean μ and SD $\sigma/\sqrt{n}$

- This means that there is a 95% chance that the (**RV**) $\overline{X}$ falls within $1.96\sigma/\sqrt{n}$ of the true population mean μ :

$$P(\mu - 1.96\sigma/\sqrt{n} \leq \overline{X} \leq \mu + 1.96\sigma/\sqrt{n}) = 0.95.$$

- Now, the RV $\overline{X}$ being within $1.96\sigma/\sqrt{n}$ of μ is the *same event* as μ being within $1.96\sigma/\sqrt{n}$ of $\overline{X}$, so the events have the same probability :

$$P(\overline{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \overline{X} + 1.96\sigma/\sqrt{n}) = 0.95.$$

CI for the population mean, cont

- The interval $(\overline{x} - 1.96\sigma/\sqrt{n}, \overline{x} + 1.96\sigma/\sqrt{n})$ based on the *(observed) sample mean* $\overline{x}$ is called a **95% confidence interval** for μ
- The value 0.95 (95%) is called the **confidence level**
- When (as is usually the case) the population SD σ is unknown, it can be estimated by the sample SD s
- Since $1.96 \approx 2$, we can express the 95% CI as : $\boxed{\overline{x} \pm 2\,\frac{s}{\sqrt{n}}}$

Example – CI (mechanics)

Eaemple 6.5 Suppose we want to estimate the mean income of a particular population. A random sample of size $n = 16$ is taken ; $\overline{x}$ = \$23,412, s = \$2000.

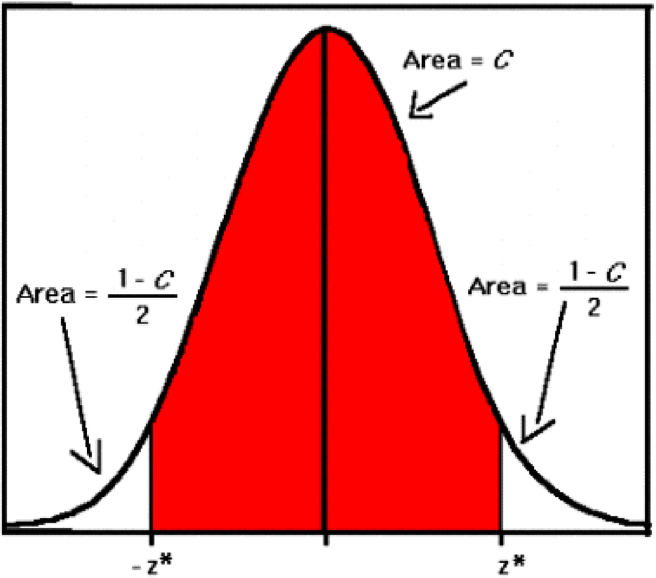
(a) Estimate the population mean μ

(b) Make an approximate 95% CI for μ **...**

Confidence level $\neq$ 95% ? ?

- The most commonly used confidence level is 95% or 90%, but there is no rule saying that we need to use this level
- The level can any value under 100%, depending on how 'confident' you want to be that the true parameter value will be contained in an interval made according to the procedure outlined above
- When the confidence level changes *the associated z-value (1.96 for a 95% CE) needs to be changed as well*

Illustration



Another example – CI mechanics

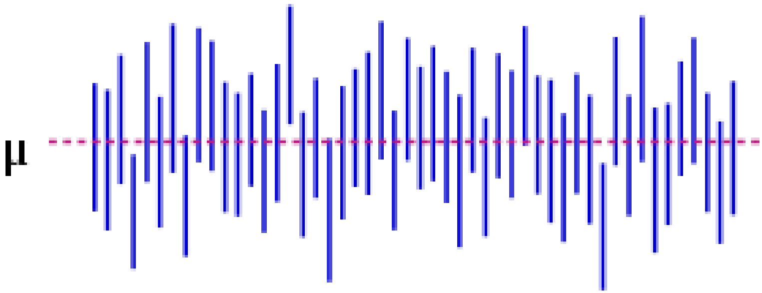
Example 6.6 Suppose we want to estimate $\mu =$ the average exam note in a arge population. A random sample of size 25 is obtained ; $\overline{x} = 72$, $s = 15$.

Give an approximate 90% CI for μ.

CI Interpretation

- It is tempting **BUT WRONG ! ! ! ! !** to believe that for a specific 95% CI there is a 95% chance that the true parameter value is in the CI – *long-run frequency interpretation of probability*
- With this interpretation, the population parameter is **NOT** a RV, but rather a **constant** whose value is unknown
- *Before* sampling, the sample mean $\overline{X}$ is a RV
- *After* sampling, **there is no longer a random variable**
- The parameter is either *in* or *out of* this particular interval
- The 95% says something about the *sampling procedure* :
  If we did the whole procedure over and over again (getting a random sample and making a 95% CI), **about 95% of the intervals made in this way would contain the true parameter value**

Illustration



Another example

Example 6.7 In a particular year there are 100,000 army recruits. The average weight is 75 kilos, with an SD of 15 kilos.

(a) If possible and appropriate, make a 95% CI for the average weight of army recruits in that year. *Explain.*

(b) Suppose now that the population mean weight is unknown, but a random sample of 400 is taken, and the average weight in the sample is 75 kilos with an SD of 15 kilos. Can you make a CI now ?

(c) Do you need to assume that the distribution of weights of army recruits is normal ? *Explain.*

CI – Suppositions

1. There is an *unknown* population parameter
2. There is a *random sample* (independent observations or SRS from a large population, where the sample size is small compared to the population size)
3. We can apply the *CLT*