Unit 3

Probability Distributions

ESIGELEC

Instructor: Federico Perea

Contents

Introduction

Probability distributions. Cumulative distribution function

Discrete distributions. Probability mass function

Measures characterizing a discrete distribution

Binomial distribution

Sampling inspection plans

Continuous distribution. Probability density function.

The normal distribution

Critical values

Summary

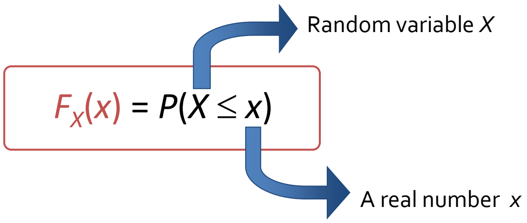
- This unit introduces basic concepts that will later allow you to calculate probabilities related with a number of different random situations.
- More precisely, the concepts of cumulative distribution function and probability mass function for discrete random variables will be introduced.
- Besides, the concepts of expected value, variance, etc. will be presented.
- Some examples of discrete random variables, Uniform, Binomial.
- To finish the unit, you will see an application of these distributions: the sampling inspection plans.

Introduction

- In unit 1 you learned how to describe a data set, constituted by observations of one or more random variables, taken from a given sample.
- You are interested in the behaviour of a random variable (how it changes, which values it can take, etc.), not only on the elements of the sample, but on the whole population under study.
- In this unit you will learn the formalisms that model the behaviour of a random variable, by means of probability distributions.
- In the discrete case, you will use the probability mass function. In the continuous case you will use the probability density function.
- You will also see some concepts associated with random variables: expected value, median, and standard deviation.
- One discrete random variable will be studied in more detail: binomial distribution.
- It models many real situations.
- You will study their main properties and in which contexts they can be applied.
- You will see one application of these distributions: sample inspection plans.

Probability distributions. Cumulative distribution function

- Given a random variable X, our interest focuses on how it changes, that is, what is the probability that the elements in the population under study take its different values.
- Cumulative distribution function of X : It is defined over any real number x:

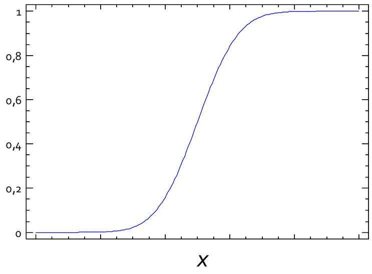


$$\boxed{F_X(x) = P(X \le x)}$$

Random variable X

A real number x

- The cumulative distribution function (c.d.f) F_X of a random variable X satisfies:
  - $F(-\infty) = \lim_{x\to-\infty} F_X(x) = P(X \le -\infty) = 0$
  - $F(+\infty) = \lim_{x\to+\infty} F_X(x) = P(X \le +\infty) = 1$
  - F is not decreasing (equivalently, monotone increasing)



- Once you know the c.d.f. F_X you can compute the probability associated to X:

$$\boxed{P(a < X \leq b) = F_X(b) - F_X(a)}$$

- Therefore, if you know F_X the behavior of X is characterized.
- Random variables can also be characterized by means of their probability mass functions, (discrete case, next section), or probability density function (continuous case).

Discrete distributions. Probability mass function

- The probability mass function (p.m.f.) of the discrete variable X is

F(X)=0, 1/6, 1/3, 1/2, ...,1

$$\boxed{p_X(x) = P(X = x)}$$

- **Example :** Consider the r.v. X "number when rolling a dice". Then:

  $p_X(1) = 1/6$, $p_X(2) = 1/6$, ... , $p_X(6) = 1/6$

  For a better understanding, draw p_X and F_X for this example.

- **Exercise :** Consider the discrete r.v. X "number of heads when flipping 3 coins". What is its probability mass function? Draw both the p.m.f. and the c.d.f.
  Solution: p(0)=p(3)=1/8; p(1)=p(2)=3/8.

- REMARK: The probabilities of all possible outcomes sum up to 1!!! That is:

$$\boxed{\sum p_X(x_i) = 1}$$

- Relationship between p_X and F_X:

$$\boxed{F_X(x) = \sum_{x_i \le x} p_X(x_i)}$$

» 9

Measures characterizing a discrete random variable

- The concept of expected value generalizes the mean of a data set (unit 1).
- Expected value of discrete random variables : The expectation, expected value or mean value of a discrete r.v. Is:

$$\boxed{\mu_X = E(X) = \sum x_i p_X(x_i)}$$

0*1/8+1*3/8+2*3/8+3*1/8

=1.5

- That is, you assign to each possible value x_i a "weight", which coincides with its probability $p_X(x_i)$, so that μ_X is the equilibrium point, the point of "weighted share" (the same idea you used when working with sample data).

» 10

- Population variance and standard deviation : Just like you did about the mean, the concept of sample variance and standard deviation can be generalized to the population (unit 1).
- Population variance : how far the observations from the mean on average are

$$\boxed{\sigma^2_X = \mathrm{Var}(X) = \mathrm{E}(X-\mu_X)^2=\mathrm{E}(X^2)-(\mu_X)^2}$$

(0-1.5)^2*1/8+(1-1.5)^2*3/8+...

- Population standard deviation :

$$\sigma_X = \sqrt{Var(X)}$$

- Just like with samples, both parameters measure the variability of the random variable (in the population).

Instead of the variance, it is more common to use its square root, because it is expressed in the same units as the data.

» 11

Binomial distribution

- The binomial distribution arises when you want to study the number of times that a given event occurs (also called success) when repeating several times a random experience with two possible outcomes. As a real application, it can be used to study the number of "defective units" in a production line.
- Two conditions must hold:
  - The probability of the event to occur is constant for all the repetitions of the experiment
  - The different repetitions of the experiment must be independent between one another.

- **Example :** What is the probability of getting 2 ones when rolling a dice 4 times? And 4 ones? And no *one* at all?

$P(X=0) = \overset{3}{(0)} *(1/2) *(1/2)\text{^}(3-0)$

» 12

- Binomial distribution: its probability mass function is:

$$\boxed{P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}}$$

  - n : Number of repetitions of the random experiment
  - p : Probability of of getting a success in one of the repetitions

- Notation :

$$\boxed{X \sim Bi(n,p)} \quad \text{or} \quad \boxed{X \sim B(n,p)}$$

» 13

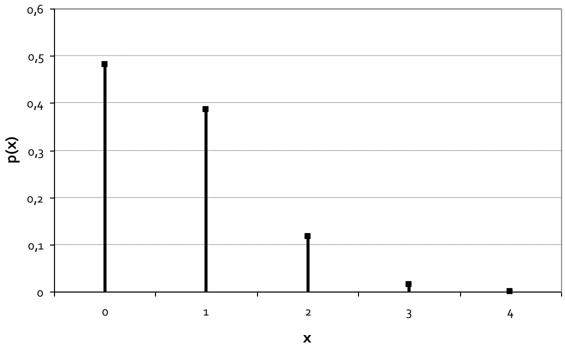
- REMARK: By using combinatorics you can prove that the p.m.f. defined before actually satisfies that: $\sum p_X(x_i) = 1$.

- **Examples :** The following r.v. follow a binomial distribution. Specify the values of n and p:
  - Number of defective screws in a 10-unit box, knowing that 10% of the screws produced at the factory are defective. n=10, p=0.1
  - Number of times that a given mechanism works when it is activated 15 times, knowing that the probability that such mechanism works when it is activated is 0.95. n=15, p=0.95
  - Number of heads when flipping a coin four times. n=4, p=1/2
  - Number of defective pages when photocopying a 100 page document, if the copy machine makes defective copies one out of 25 times. n=100, p=1/25

» 14

3.5

- **Examples :** Bi($n = 4$, $p = 1/6$)



» 15

- The expected value and variance of a binomial distribution X with parameters n and p:

$$X \sim \mathrm{Bi}(n , p) \quad \Rightarrow$$

$$\boxed{\mathrm{E}[X] = n \cdot p} \quad \text{and} \quad \boxed{\mathrm{Var}[X] = n \cdot p \cdot (1-p)}$$

- **Exercise :** How many times can you expect to get *one*, on average, when rolling a dice 6 times? Solution: 1.

6*1/6

» 16

- **Exercise :** Due to a problem in its production line, a factory produces 2% of defective light bulbs. These light bulbs are sold in 10-unit boxes. 40% of the boxes are red, the rest are blue. If you randomly pick one box, what is the probability that such box has more than two defective light bulbs?
  Solution: 1-*pbinom(2, size=10, prob=0.02)* = 0.00086. pbinom=P[X<=a]

0.1

- **Exercise :** Certain banner of a website shows, on average, one out of 10 times you visit such website. What is the probability that the banner appears within the next five visits to this site? If you know that the banner did appear last time you visited this site, what is the value of the previous probability?
  Solution: *1-pbinom(0, size=5, prob=0.1)* = *1-dbinom(0, size=5, prob=0.1)* = 0.409.

dbinom=P[x=a]

- **Exercise :** A company has put a small present in one out 100 chip bags. If you buy 200 bags, what is the expected number of presents you are to get? What is the probability that you do NOT find any present?
  Solution: 2, *dbinom(0, size=200, prob=0.01)* = *pbinom(0, size=200, prob=0.01)* = 0.13398

- Computation : In order to compute probabilities related with a binomial r.v. you have its probability mass function. But in some cases this computation is difficult (or even impossible), specially when n is large.
- Approximations :
  - When *n* is sufficiently large and *p* is very small or very large, you can approximate the binomial by a Poisson distribution (not included in this course).
  - When *n* is sufficiently large and *p* is close to 0.5 you can approximate the binomial by a normal distribution (to be seen later).
- There are many other discrete probability distributions: *uniform, poisson,…*

Sampling inspection plans

- Sampling inspection plan : Procedure to decide whether or not a certain batch is accepted, based on inspecting a random sample of this batch. More precisely, you will base your decision on the number of defective units found in the sample

large groupe

- BATCH (“infinite” size) ⇒ random sample (size n)
- Random variable of interest :

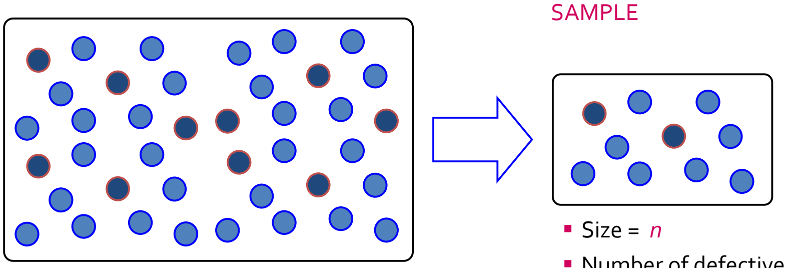
X = number of defective units in the sample

refuse or accept

- Simple sampling inspection plan. It is the most common procedure :
  - If $X \leq c$, accept the batch.
  - Si $X > c$, reject the batch.
- Therefore, parameter c represents the maximum admissible number of defective pieces in the sample to accept the batch.
- Assuming in the batch the probability that a unit is defective is constant and equal to p, then you have that:

$$X \sim \text{Bi}(n\ ,\ p)$$

BATCH



SAMPLE

- Size = "∞"
- Actual (but "unknown") proportion of defective units = p

- Size = n
- Number of defective units observed = X (observed proportion of defective units = X/n)

SIMPLE SAMPLING INSPECTION PLAN:

- IF $X \leq c$ ⇒ Accept BATCH
- If $X > c$ ⇒ Reject BATCH

- If $X \sim \text{Bi}(n, p)$, you can compute the probabilities of accepting or rejecting batches as if it was a binomial distribution problem .
- But in these problems the unknown is n or c (or both!). A typical question is to determine the sampling inspection plan (that is, compute n and c) in such a way that the probability of accepting a "bad" batch is low.
- Summary:
  - n : Sample size.
  - X : Random variable of interest: number of defective units in sample.
  - c : Maximum admissible number of defective units in our sample (if there are more, the whole batch is rejected).
  - p : proportion of defective units in the batch.
  - R : probability of rejecting the batch.

- **Exercise :** A company produces plastic pieces for planes. One of its best clients demands batches in which less than 5% of pieces are defective. The company wants to have a sampling inspection plan over each of the batches to be sent to this client. Therefore, a sample of size *n* will be analyzed and, in case 2 or more defective pieces are found, the whole batch will be rejected.

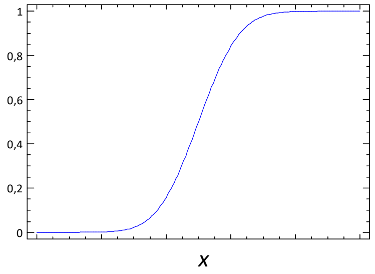
  Find the minimum value of *n* so that a batch considered “not acceptable” by the client is rejected with probability 0,9 or more.

  Solution: n=77.

Continuous distributions

A range of probability

- The cumulative distribution function (CDF) F_X of a random variable X satisfies:
  - $F(-\infty) = P(X \leq -\infty) = 0$
  - $F(+\infty) = P(X \leq +\infty) = 1$
  - F is not decreasing (equivalently, monotone increasing)

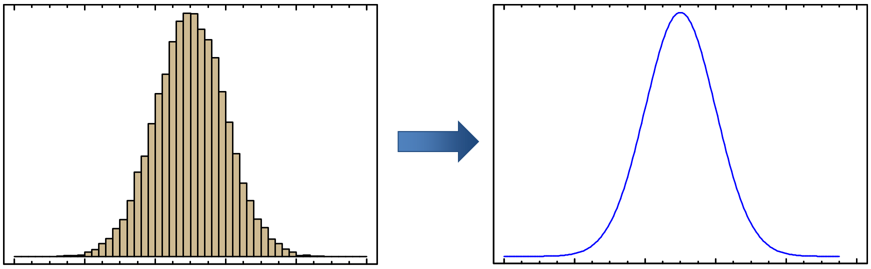


if X is continuous, P(X=a)=0

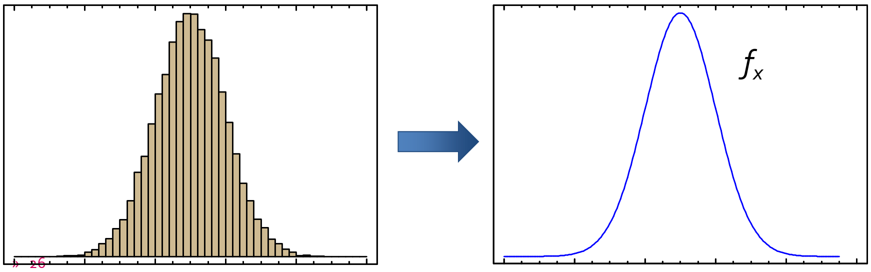
F(a)=P(X<=a) is a continuous function

f(a) -> Probability density function

- Probabilities are a generalization of relative frequencies.
- Let X be a continuous random distribution. If you had a large enough sample, you could represent it by means of a histogram with many intervals. In fact, if you had the complete population (theoretically infinite), you could represent it (theoretically) with a histogram with infinite many intervals.



- This function (the limit of the histogram when you increase the number of intervals) is the probability density function of the random variable X, and is represented by f_X
- Remember that the concept of probability, which is applied over a population, is a generalization of the concept of relative frequency over a sample.



for large sample

for population

Proportion of observations between *a* and *b*:

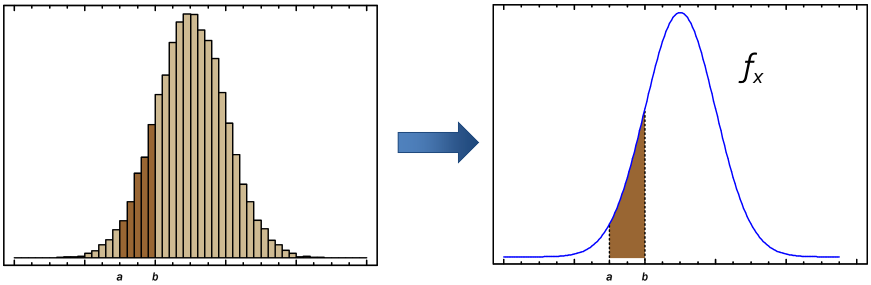
$f_{[a,b]}$ = darker area

Probability between *a* and *b*:

$P(a < X \leq b)$ = darker area

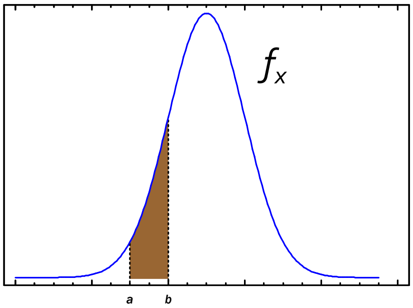
Sample

Population



- Therefore, calculating probabilities reduces to calculating areas. The area between the curve and the OX axis is the following integral

$$\boxed{P(a < X \leq b) = \int_a^b f_X(x)\,dx}$$



- Relation between f_X and F_X:

$$\boxed{F(x) = \int_{-\infty}^{x} f_X(t)dt} \qquad \boxed{f(x) = F'(x)}$$

- Expected value for continuous random variables: Instead of summing like in the discrete case, you now integrate:

$$\boxed{\mu_X = E(X) = \int x f_X(x)dx}$$

- The idea behind this formula is the same, the "equal share" point.

- **Exercise :** Given the following function, associated to the random variable *X*,

$$f(x) = \begin{cases} \frac{1}{2}(x-1)^3 + \frac{1}{2}\ , & 0 \le x \le 2 \\ 0\ , & \text{otherwise} \end{cases}$$

a) Check that this is a probability density function.

b) Compute the probability that X is less than 1'8.

c) Compute the expected value of *X*.

d) Compute the variance of *X*.

Solution : b) 0,8262 c) 6/5 d) 22/75

- **Exercise :** Compute the 1st quartile of *X* knowing that its probability density function is $f_X(x) = 4x^3$, $0 \le x \le 1$ (and 0 otherwise).

Solution : 0,7071

Normal distribution

Normal Bell-shaped Gaussian

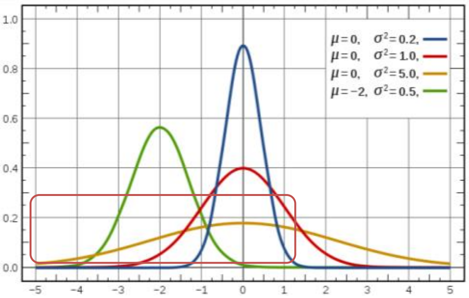
- The normal distribution is one of the most common continuous distributions because it models lots of random variables.
- It is a symmetric model: the probability density focuses on the central area, and symmetrically decreases as you go far from the central value. It is bell shaped.
- The normal distribution is not the only bell-shaped random distribution, but it is the one that most often arises when observing and studying random behaviour in natural sciences, social sciences, etc.
- If you have a sample, you may check if these data come from a normal population or not: if both Kurtosis and Assimmetry coefficients are close to zero (e.g. in the range [-2,2]), you may accept that these data are normally distributed.
- There are other tests to check normality: normal probability plot, Shapiro-Wilk test, Kolmororov-Smirnov test,…

If qqnorm() shows aligned points or

shapiro.test() p-values > 0.05 the simple can be considered as normal

- Normal distribution : Characterized by the following probability density function *f*:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- μ : Expected value
  Location parameter.
- σ : Standard deviation
  Dispersion parameter.

$F(x) = P(X \leq x)$ NO analytical expression Notation: $X \sim N(\mu, \sigma)$

In R: $pnorm(x, mean = \mu, sd = \sigma, lower.tail = TRUE)$

- Mean and variance:

$$\boxed{X \sim \mathrm{N}(\mu , \sigma) \quad \Rightarrow \quad \mathrm{E}[X] = \mu \quad \text{and} \quad \mathrm{Var}[X] = \sigma^2}$$

- In order to compute probabilities, you need a table or a software.
- If $X \sim \mathrm{N}(\mu , \sigma)$, then:
  - $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0{,}6827 \approx 2/3$
  - $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0{,}9545 \approx 95\%$
  - $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0{,}9973 \approx 99{,}7\%$

Exemple: P(X<=4.9) = pnorm(4.9, mean=5, sd=0.1, lower tail=TRUE)

- **Exercise :** The weight of a box of circuits manufactured in a certain factory is normally distributed with mean 5 kg and standard deviation 0'1 kg. In what interval is the weight of the 95% of boxes produced in this factory expected to be? SOLUTION: (4.8,5.2)

P(X>4.8) = pnorm(4.8, mean=5, sd=0.1, lower tail=FALSE)

- **Exercise :** some 99'7% of people in a certain population are between 1'51 m and 1'93 m tall. The mean height in this population is 1'72 m. Assuming that this variable follows a normal distribution, estimate the standard deviation of such variable. SOLUTION: $\sigma = 0.07$.

Other distributions to be used and critical points

- Chi-squared distribution
  - Sum of standard normal distributions squared
  - $\chi_k^2 = \sum_{i=1}^{k} Z_i^2$, k are the *degrees of freedom*
- Student's t distribution
  - Ratio between a standard normal distribution and a chi-squared distribution divided by its degrees of freedom, both independent from each other
  - $T_k = Z \,/ \sqrt{\frac{\chi_k^2}{k}}$, k are the degrees of freedom
- Snedecor's F distribution
  - Ratio between two chi-squared divided by their degrees of freedom
  - $F_{k_1,k_2} = \frac{\chi_{k_1}^2 \,/\, k_1}{\chi_{k_2}^2 \,/\, k_2}$, k_1, k_2 are the degrees of freedom of the numerator and the denominator, respectively.

Notation and other random variables: critical points

- The sample mean (known): $\bar{x}$
- The sample standard deviation (known): S
- The population mean (unknown): μ
- The population standard deviation (unknown): σ
- The critical value of a t-student distribution: t_n^{σ}
- The critical value of a 0-1 normal distribution: z_{α}
- The critical value of a chi-square distribution: $\chi_{n,\alpha}^2$
- The critical value of a Snedecor's F-distribution: F_{k_1,k_2}^{α}
- These critical points can be obtained from tables or from statistical software (for example, Microsoft Excel).

Critical values

$$\alpha \in [0,1]$$

$$z_\alpha : P[Z > z_\alpha] = \alpha$$

$$t_n^\alpha : P[T_n > t_n^\alpha] = \alpha$$

$$F_{k_1,k_2}^\alpha : P[F_{k_1,k_2} > F_{k_1,k_2}^\alpha] = \alpha$$

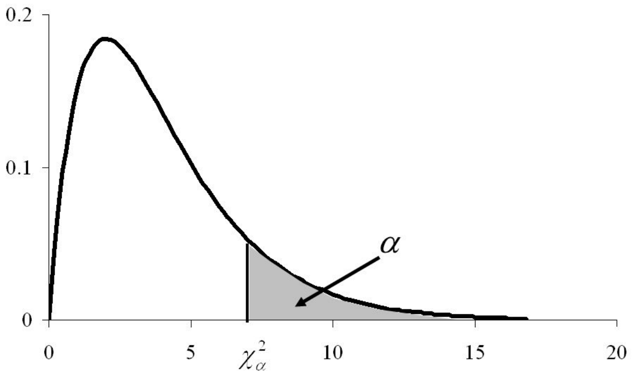
$$\chi_k^{2,\alpha} : \mathrm{P}[\chi_k^{2} > \chi_k^{2,\alpha}] = \alpha$$

Example of critical point

Find a point so that P(X>=a)=alpha,,,   standard alpha =0.05

, a is a critical point

Critical Points of the χ^2 distribution



In R…

qnorm(alpha, mean=miu, sd=cigma, lower tail=FALES), by default, miu=0, cigma=1

$Z_\alpha = qnorm(\alpha, lower.tail = FALSE)$

$t_n^\alpha = qt(\alpha, n, lower.tail = FALSE)$

$F_{k_1,k_2}^\alpha = qf(\alpha, k_1, k_2, lower.tail = FALSE)$

$\chi_k^{2,\alpha} = qchisq(\alpha, k, lower.tail = FALSE)$

$Z_{0.05} = qnorm(0.05, lower.tail = FALSE) =$ qnorm(0.95, lower.tail=TRUE) = 1.644

$t_{15}^{0.05} = qt(0.05, 15, lower.tail\ =\ FALSE)$ = 1.75305

$F_{9,4}^{0.05} = qf(0.05, 9, 4, lower.tail\ =\ FALSE) = 5.998779$

$qchisq(0.025, 8, lower.tail\ =\ FALSE) =\ 17.53455$

After this…

- You are ready to estimate population parameters. But, what is this????
- estimate: “roughly calculate or judge the value, number, quantity, or extent of something.”
- A population may consist of too many elements: millions!
- You want to compute the average of a population.
- What you do is:
  - Take a proper sample from the population (not too many)
  - Analyze the elements of the sample, and compute the average
  - With the techniques you will see in the next section, you will approximate (estimate) the average of the whole population
- This will save you time and money!