Review of Probability

First Version - August 29, 2016
Present Version - September 5, 2020

Instructors

- Angelo Secchi
    - angelo.secchi@univ-paris1.fr
    - in case you need to talk with me, drop an email to make an appointment.
    - Room R4-70, 4th floor. Paris School of Economics, 48 Boulevard Jourdan, 75014.

- Nina Rapoport
    - nina.rapoportpsemail.eu
    - tbc

Instructors

- Angelo Secchi, will get you lost;
- Nina Rapoport, will come to rescue you.

Organization

- Syllabus [here]
- Slides available on my personal page [here]
  - warnings on slides: howto use them, handle them with care, first draft
- Exams (boot camp + Introduction to Econometrics):
  - Mid-term (70%bc + 35%imetrics): Thursday Nov 5, 09h30. (Room, TBC)
  - Take home (30%bc + 15%imetrics): deadline Dec 17, after class.
  - Final (50%imetrics): TBC. (Room, TBC)

Organization

- (Hopefully useless) remarks:
  - Exams will not be rescheduled. No exceptions.
  - In case of exceptional events contact, as soon as possible, the Director of the Master.
  - Be on time.
  - Class is a mobile-free zone.

- Typical students attitude:
  - Psycho: "I do not understand a single word; no problem I will catch-up later on during the semester";
  - Econ: "I already know Econometrics; no need to properly follow classes and to work hard".

Organization

- (Hopefully useless) remarks:
  - Exams will not be rescheduled. No exceptions.
  - In case of exceptional events contact, as soon as possible, the Director of the Master.
  - Be on time.
  - Class is a mobile-free zone.

- Typical students attitude:
  - Psycho: "I do not understand a single word; no problem I will catch-up later on during the semester";
  - Econ: "I already know Econometrics; no need to properly follow classes and to work hard".

Both generate dramatic failures in terms of learning and grades!

Organization

- (Hopefully useless) remarks:
  - Exams will not be rescheduled. No exceptions.
  - In case of exceptional events contact, as soon as possible, the Director of the Master.
  - Be on time.
  - Class is a mobile-free zone.
- Typical students attitude:
  - Psycho: "I do not understand a single word; no problem I will catch-up later on during the semester";
  - Econ: "I already know Econometrics; no need to properly follow classes and to work hard".

  Both generate dramatic failures in terms of learning and grades!
- ... and yes life is typically unfair. For Psycho students my classes will be on average harder. But last 3 years average grades were almost identical.

R

- R is a statistical programming language:
  - it is free and open source, and always will be
  - it is a programming language rather than a graphical interface, it allows scripting
  - it has a very active and helpful online community

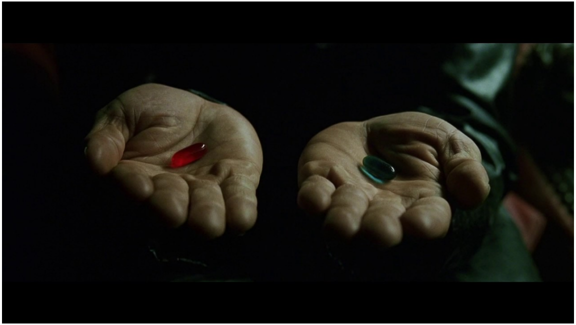
- R does not have a graphical user interface, RStudio is a graphical front-end to R

- A simple starting point for newbies: https://ourcodingclub.github.io/tutorials/intro-to-r/

- A more complete Intro: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

what would you prefer?

- 🟥 Knowledge, freedom, uncertainty and the brutal truths of reality
- 🟦 Security, happiness, beauty, and the blissful ignorance of illusion



what would you prefer?

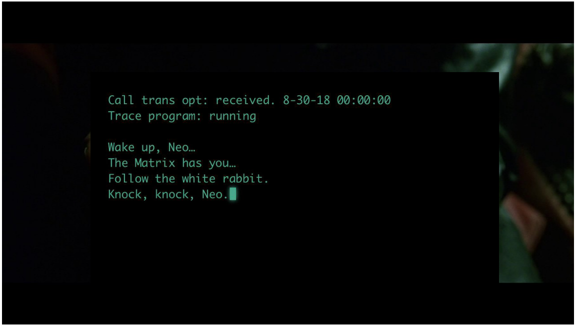
🔴 Knowledge, freedom, uncertainty and the brutal truths of reality

🔵 Security, happiness, beauty, and the blissful ignorance of illusion

what would you prefer?

🟥 Knowledge, freedom, uncertainty and the brutal truths of reality

🟦 Security, happiness, beauty, and the blissful ignorance of illusion



```
Call trans opt: received. 8-30-18 00:00:00
Trace program: running

Wake up, Neo…
The Matrix has you…
Follow the white rabbit.
Knock, knock, Neo.█
```

The Matrix. Directed by The Wachowskis, 1999.

Section 1

One random variable

Random variables

Many events have an element of chance or randomness in the sense that there is something not yet known that is eventually revealed. For these events:

- the mutually exclusive potential results are called outcomes;
- the probability of an outcome is the proportion of the time that the outcome occurs in the long run;
- the set of all possible outcomes is called sample space;
- an event is a subset of the random space that is a set of one or more outcomes.

- "tossing a fair coin":

  1(head) 0(tails) [outcomes]

  0.5(head) 0.5(tails) [probability of each outcome]

  {1,0} [sample space]

  "not observing 1" [an event]

- "tossing a fair coin":

  1(head) 0(tails) [outcomes]

  0.5(head) 0.5(tails) [probability of each outcome]

  {1, 0} [sample space]

  "not observing 1" [an event]

- "rolling a fair dice":

  3 2 1 6 4 5 [outcomes]

  1/6(3) 1/6(2) 1/6(1) 1/6(6) 1/6(4) 1/6(5)

  [probability of each outcome]

  {3, 2, 1, 6, 4, 5} [sample space]

  "observing $\geq$ 4" [an event]

Discrete random variables

We define the discrete random variable X a variable whose value is determined by the outcome of a chance experiment.

If X represents "tossing a fair coin", then

1(head) 0(tails) [outcomes]

0.5(head) 0.5(tails) [probability of each outcome]

{1,0} [sample space]

Discrete random variables

We define the discrete random variable X a variable whose value is determined by the outcome of a chance experiment.

If X represents "tossing a fair coin", then

$x_1 = 1 \quad x_2 = 0$ [2 possible outcomes]

$\Pr(X = x_1) = p_1 = 0.5 \quad \Pr(X = x_2) = p_2 = 0.5$ [probability of each outcome]

$\{x_1, x_2\}$ [sample space]

In general a random variable X is described as

$x_1, x_2 \ldots, x_i, \ldots x_k$ [k possible outcomes]

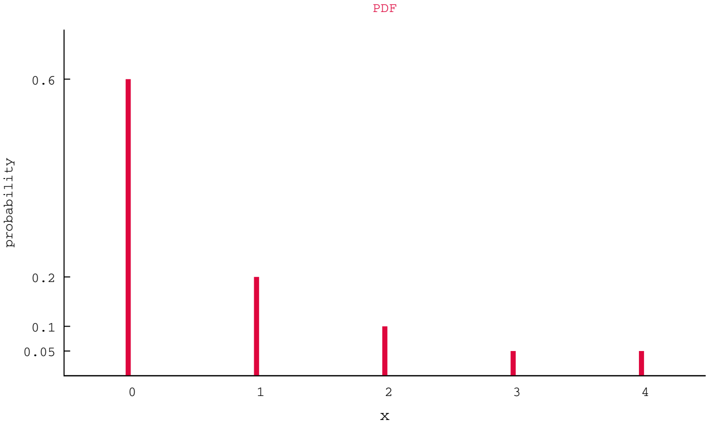
$p_1, p_2 \ldots, p_i, \ldots p_k$ [probability of each outcome]

$\{x_1, x_2, \ldots, x_i, \ldots x_k\}$ [sample space]

where $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$.

Probability Distribution Function - PDF

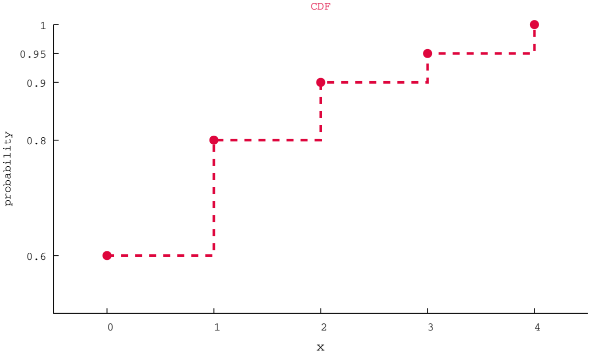
For any discrete random variable X one can define its probability distribution function, PDF(x) as the list of all possible values of the variable with the probability that each value will occur.



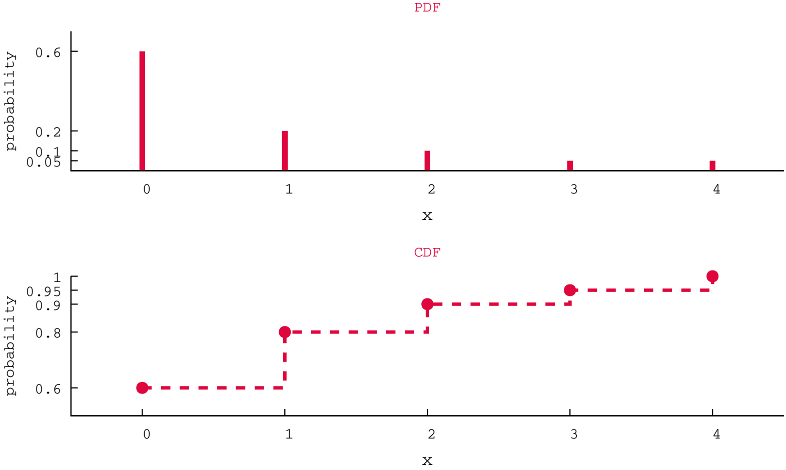
Cumulative Distribution Function - CDF

For any discrete random variable X one can define its cumulative distribution function, CDF(x), as the probability that the random variable is less than or equal to a particular value. Formally, CDF(x)=Pr(X≤x), where

- CDF(x) is obtained by summing the pdf over all values $x_i \leq x$;
- CDF(x) is a non-decreasing function of x.



PDF and CDF of a discrete random variable



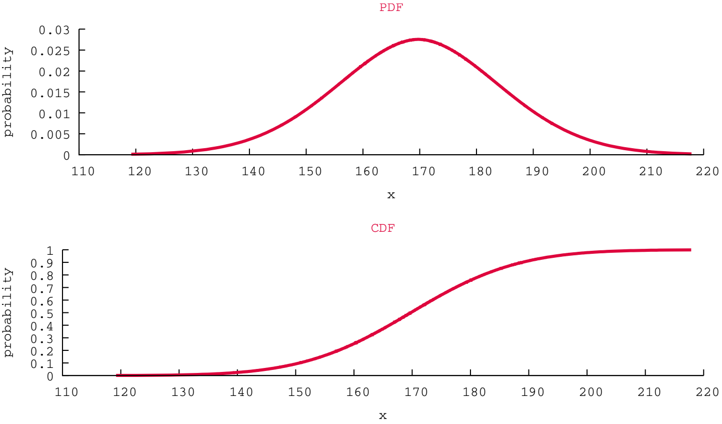
Continuous random variables

For a continuous random variable all the definitions remain (more or less) the same. However, since a continuous random variable can take on a continuum of possible values:

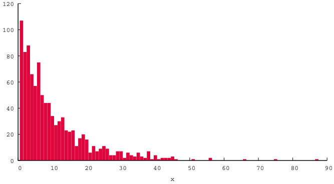
- we cannot list anymore each possible value of the random variable. Instead we define the probability density function (PDF);

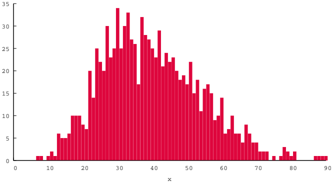
- the area under the PDF between any two points is the probability that the random variable falls between these two points.

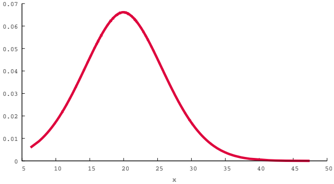
X is the random variable representing "the height (in cm) of a human being".

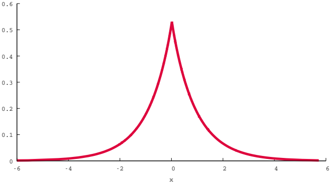


Examples


Geometric


Negative binomial


Chi-squared


Laplace

Practice

Exercise 1. Consider the random experiment of tossing simultaneously two regular coins. First,

- list all the possible outcomes of this experiment.

Assume that all these outcomes have the same probability of being observed and let Y denote the number of "heads" obtained. Compute

- the probability distribution function of Y;
- the cumulative distribution function of Y.
- graph the CDF.

Moments of a distribution

Often we need to synthetically describe the shape of a PDF.

In order to do that we introduce measures capturing different aspects of the shape:

- the mean (or expected value) which locates the distribution;
- the standard deviation (or spread) which measures its variability;
- the skewness which measures the lack of symmetry of the distribution;
- the kurtosis which measures how thick, "fat", or heavy are its tails. [the likelihood of observing realizations far away from the mean]

Moments of a distribution

Often we need to synthetically describe the shape of a PDF.

In order to do that we introduce measures capturing different aspects of the shape:

- the mean (or expected value) which locates the distribution;
- the standard deviation (or spread) which measures its variability;
- the skewness which measures the lack of symmetry of the distribution;
- the kurtosis which measures how thick, "fat", or heavy are its tails. [the likelihood of observing realizations far away from the mean]

Mean

The mean of a random variable X, denoted E(X), is the long run average value of the rv over many repeated trials or occurrences.

Mean

The mean of a random variable X, denoted E(X), is the long run average value of the rv over many repeated trials or occurrences.

- X represents "tossing a fair coin":

$x_1 = 1 \quad x_2 = 0$ [possible realizations of X, k=2]

$p_1 = 0.5 \quad p_2 = 0.5$ [probability of each realization]

$$E(X) = x_1 p_1 + x_2 p_2 = 1 * 0.5 + 0 * 0.5 = 0.5 \quad .$$

Mean

The mean of a random variable X, denoted E(X), is the long run average value of the rv over many repeated trials or occurrences.

- Y represents "rolling a fair dice":

$$x_1 = 3 \;\; x_2 = 2 \;\; x_3 = 1 \;\; x_4 = 6 \;\; x_5 = 4 \;\; x_6 = 5$$

$$p_1 = 1/6 \;\; p_2 = 1/6 \;\; p_3 = 1/6 \;\; p_4 = 1/6 \;\; p_5 = 1/6 \;\; p_6 = 1/6$$

$$E(X) = x_1p_1 + x_2p_2 + x_3p_3 + x_4p_4 + x_5p_5 + x_6p_6 = 3.5$$

Mean

The mean of a random variable X, denoted E(X), is the long run average value of the rv over many repeated trials or occurrences.

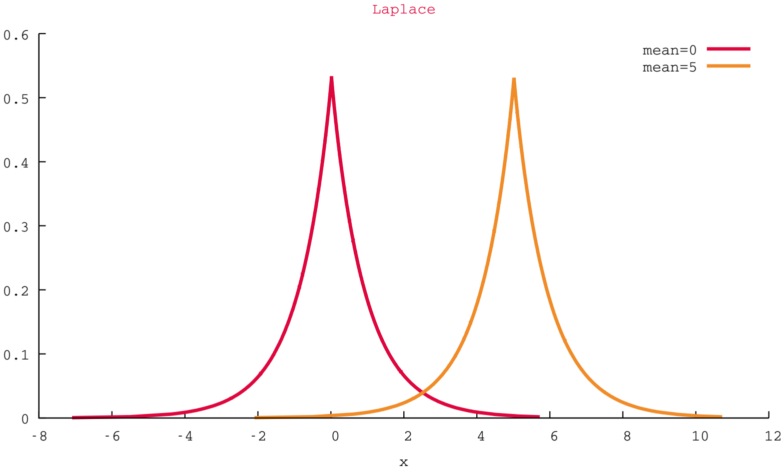
In general, for a discrete rv the mean is defined as

$$E(X) = \mu_X = \sum_{i=1}^{k} x_i p_i \quad ,$$

where x_i are all the k outcomes of the rv X and p_i the corresponding probabilities.

Remark. The mean is then a weighted average of the possible realizations of a rv where the weights are the true probabilities of each realization.

Mean



a location shift

Other location parameters

Often to locate a distribution we also consider

- MEDIAN(X) defined as any real number m_X that satisfy

$$Pr(X \leq m_X) \geq 0.5 \text{ and } Pr(X \geq m_X) \geq 0.5 \quad ;$$

- MODE(X) as the value of X with the highest probability of being drawn.

In case of a symmetric uni-modal distribution mean, median and mode coincide.

Variance

The variance of a random variable X, denoted var(X) or σ_X^2, measures the dispersion or the "spread" of a probability distribution. For a discrete rv

$$\text{var}(X) = \sum_{i=1}^{k} (x_i - \mu_x)^2 p_i \quad ,$$

where x_i are the outcomes of the rv, p_i the corresponding probabilities and μ_X the mean of X.

Variance

The variance of a random variable X, denoted var(X) or σ_X^2, measures the dispersion or the "spread" of a probability distribution. For a discrete rv

$$\mathrm{var}(X) = \sum_{i=1}^{k} (x_i - \mu_x)^2 p_i \quad ,$$

where x_i are the outcomes of the rv, p_i the corresponding probabilities and μ_X the mean of X.

- Y represents "tossing a fair coin":

$$x_1 = 1 \;\; x_2 = 0$$

$$p_1 = 0.5 \;\; p_2 = 0.5 \qquad E(X) = \mu_X = 0.5$$

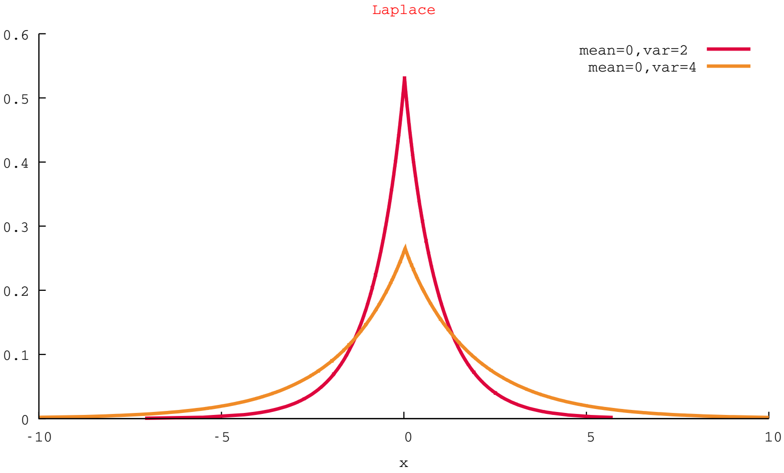
$$\mathrm{var}(X) = \sigma_X^2 = (x_1 - \mu_x)^2 p_1 + (x_2 - \mu_x)^2 p_2 = 0.25$$

Variance

Two important remarks:

- note that var(X) is by definition equal to $\mathtt{E}\left[(\mathtt{X}-\mu_{\mathtt{X}})^2\right]$
- $\sqrt{\mathtt{var(X)}}$=sd(X)=$\sigma_{\mathtt{X}}$ represents the standard deviation of X.

Variance



no location shift, higher variability

Skewness

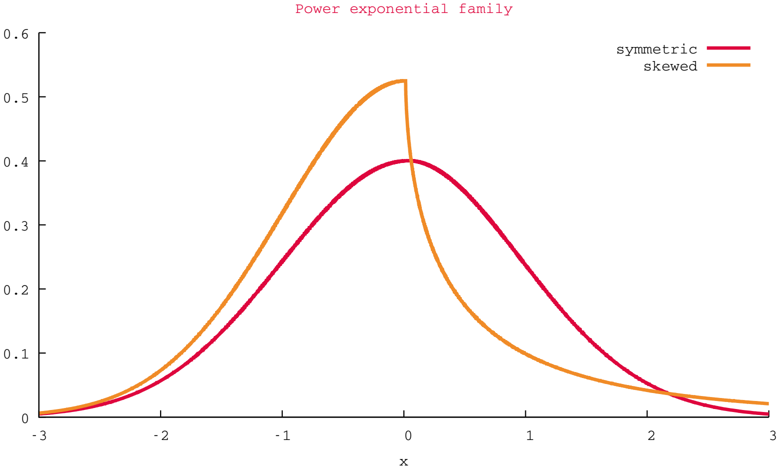
In a symmetric distribution a value of X larger than its mean of a given amount is just as likely as a value lower than its mean of the same amount.

Skewness

The skewness of a random variable X provides a mathematical way to describe how much a distribution deviates from symmetry. For a discrete rv

$$\mathtt{skewness} = \frac{\sum_{i=1}^{k}(x_i - \mu_x)^3 p_i}{\sigma_X^3} \quad .$$

Skewness



a tilt in the distribution, a probability mass shift

Kurtosis

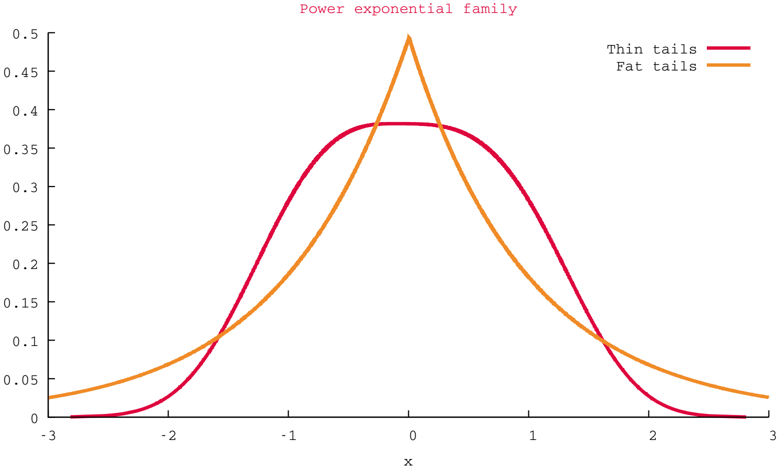
Sometimes we are interested in measuring how likely is to observe extreme observations in a distribution.

Kurtosis

The kurtosis of a random variable X provides a mathematical way to describe how much probability mass is in the tails of the distribution. For a discrete rv

$$\text{kurtosis} = \frac{\sum_{i=1}^{k} (x_i - \mu_x)^4 p_i}{\sigma_X^4} \quad .$$

Kurtosis



more peaked with fatter tails

Practice

Exercise 2. Compute the mean, variance, skewnees and kurtosis of the distribution of Y of the previous exercise.

Functions of a random variable

If X is a random variable with mean and variance $\mu_{\mathtt{X}}$ and $\sigma^2_{\mathtt{X}}$ respectively and Y=a+bX, then

$$\mathtt{E}(\mathtt{X}^2) = \sigma^2_{\mathtt{X}} + \mu^2_{\mathtt{X}}$$

$$\mathtt{E}(\mathtt{Y}) = \mathtt{E}(\mathtt{a} + \mathtt{bX}) = \mathtt{a} + \mathtt{b}\mu_{\mathtt{X}}$$

$$\mathtt{var}(\mathtt{Y}) = \mathtt{var}(\mathtt{a} + \mathtt{bX}) = \mathtt{b}^2\sigma^2_{\mathtt{X}} \quad .$$

Practice

Exercise 3. Let X denote the pre-tax earnings of an household. X is a rv with mean and variance μ_X and σ_X^2 respectively. Consider the following tax-scheme:

- household is taxed is 20% rate;
- household is given a tax-free grant of 2000$.

Express the after tax income Y as a function of the pre-tax income X and compute its mean and variance. Assume that both X and Y are discrete random variables.

Section 2

Two random variables

Most of the interesting questions involve two or more random variables.

Example. Raising the question

- how does Secchi's commuting time from home to work change when it is raining?

concerns the distribution of two random variables considered together: "commuting time" (rv Y) and "weather conditions" (rv X).

Joint distribution

The joint distribution of two discrete random variables, say X and Y, describes the probability that the two rv simultaneously take on certain values, say x and y and it is usually expressed as Pr(X=x,Y=y).

Joint distribution

	Rain (X=0)	No rain (X=1)
Long commute (Y=50 min)	0.15	0.07
Short commute (Y=10 min)	0.15	0.63

Joint distribution

The joint distribution of two discrete random variables, say X and Y, describes the probability that the two rv simultaneously take on certain values, say x and y and it is usually expressed as Pr(X=x,Y=y).

	Rain (X=0)	No rain (X=1)
Long commute (Y=50 min)	0.15	0.07
Short commute (Y=10 min)	0.15	0.63

Note that the probabilities inside the matrix sum to 1

$$\sum_{j=1}^{h}\sum_{i=1}^{k} \Pr(X = x_i, Y = y_j) = 1 \quad .$$

Joint distribution

The joint distribution of two discrete random variables, say X and Y, describes the probability that the two rv simultaneously take on certain values, say x and y and it is usually expressed as Pr(X=x,Y=y).

	Rain (X=0)	No rain (X=1)
Long commute (Y=50 min)	0.15	0.07
Short commute (Y=10 min)	0.15	0.63

In this example

$$\sum_{j=1}^{2}\sum_{i=1}^{2} \Pr(X = x_i, Y = y_j) = 0.15 + 0.7 + 0.15 + 0.63 = 1 \quad .$$

Marginal probability distribution

Using the joint distribution of X and Y one can obtain the individual distribution of both X and Y. In this context they are called marginal probability distributions, there are two of them.

In our example this means obtaining the distribution of "commuting time", not considering the effect of weather conditions, or of "weather conditions", irrespective of commuting time.

Marginal probability distribution

Using the joint distribution of X and Y one can obtain the individual distribution of both X and Y. In this context they are called marginal probability distributions, there are two of them.

In our example this means obtaining the distribution of "commuting time", not considering the effect of weather conditions, or of "weather conditions", irrespective of commuting time.

Marginal probability distribution

The marginal probability distribution of X and Y are computed from the joint distribution with

$$\mathtt{Pr}(X = x) = \sum_{j=1}^{h} \mathtt{Pr}(X = x, Y = y_j)$$

$$\mathtt{Pr}(Y = y) = \sum_{i=1}^{k} \mathtt{Pr}(X = x_i, Y = y).$$

Marginal probability distribution

In our example the marginal probability distribution of Y is computed as

$$\Pr(Y = 50) = \sum_{i=1}^{h} \Pr(X = x_i, Y = 50) = 0.15 + 0.07 = 0.22$$

$$\Pr(Y = 10) = \sum_{i=1}^{k} \Pr(X = x_i, Y = 10) = 0.15 + 0.63 = 0.78.$$

and it is typically represented as

	Rain (X=0)	No rain (X=1)	P(Y)
Long commute (Y=50 min)	0.15	0.07	0.22
Short commute (Y=10 min)	0.15	0.63	0.78

Conditional distribution

Starting from the joint distribution one can also obtain the distribution of Y conditional on having X equal to a pre-specified value x_i. These distributions are called conditional probability distributions.

In our example this means obtaining the distribution of "commuting time" (Y) conditional on fixing the weather conditions to "rain" (X=0). In this example we have 4 different conditional distributions.

Conditional distribution

Starting from the joint distribution one can also obtain the distribution of Y conditional on having X equal to a pre-specified value x_i. These distributions are called conditional probability distributions.

In our example this means obtaining the distribution of "commuting time" (Y) conditional on fixing the weather conditions to "rain" (X=0). In this example we have 4 different conditional distributions.

Conditional distribution

The conditional distribution of Y given X, Pr(Y=y|X=x) are computed from the joint distribution using

$$\mathtt{Pr(Y = y|X = x) = \frac{Pr(X = x, Y = y)}{Pr(X = x)}} \quad .$$

Conditional distribution

In our example the probability distribution of Y conditional on X=0 (when it is raining) is computed as

$$\mathtt{Pr}(Y = 50|X = 0) = \frac{\mathtt{Pr}(X = 0, Y = 50)}{\mathtt{Pr}(X = 0)} = \frac{0.15}{0.15 + 0.15} = 0.50$$

$$\mathtt{Pr}(Y = 10|X = 0) = \frac{\mathtt{Pr}(X = 0, Y = 10)}{\mathtt{Pr}(X = 0)} = \frac{0.15}{0.15 + 0.15} = 0.50.$$

and conditional on X=1 (when it is not raining) as

$$\mathtt{Pr}(Y = 50|X = 1) = \frac{\mathtt{Pr}(X = 1, Y = 50)}{\mathtt{Pr}(X = 1)} = \frac{0.07}{0.07 + 0.63} = 0.10$$

$$\mathtt{Pr}(Y = 10|X = 1) = \frac{\mathtt{Pr}(X = 1, Y = 10)}{\mathtt{Pr}(X = 1)} = \frac{0.63}{0.07 + 0.63} = 0.50.$$

Conditional expectation

Using the conditional distribution we can define the conditional expectation of Y given X that is simply the mean of the conditional distribution of Y given X

$$\mathtt{E}(\mathtt{Y}|\mathtt{X}=\mathtt{x})=\sum_{\mathtt{i}}\mathtt{y}_{\mathtt{i}}\mathtt{Pr}(\mathtt{Y}=\mathtt{y}_{\mathtt{i}}|\mathtt{X}=\mathtt{x})\quad .$$

Conditional expectation

What is Secchi's expected commuting time when it is raining?

$$
\begin{aligned}
E(Y|X=0) &= \sum_{i=1}^{2} y_i Pr(Y=y_i|X=0) = \\
&= y_1 Pr(Y=y_1|X=0) + y_2 Pr(Y=y_2|X=0) = \\
&= 50 * Pr(Y=50|X=0) + 10 * Pr(Y=10|X=0) = \\
&= 50 * 0.5 + 10 * 0.5 = 30 \quad .
\end{aligned}
$$

Conditional expectation

Using the conditional distribution we can define the conditional expectation of Y given X that is simply the mean of the conditional distribution of Y given X

$$E(Y|X = x) = \sum_i y_i Pr(Y = y_i|X = x) \quad .$$

Remarks.

- Conditional expectation is a key concept for regression.
- Similarly one can define conditional variance, conditional skewness and conditional kurtosis.

Practice

Exercise 4. Using the commuting time example compute:

- mean of X and Y;
- mean of Y conditional on X=0 and X=1, that is E[Y|X=0] and E[Y|X=1];
- the weighted mean of E[Y|X=0] and E[Y|X=1], using the probabilities of X=0 and X=1 respectively.

Law of iterated expectations

The law of iterated expectations states that the mean of Y is the weighted average of the conditional expectations of Y given X, weighted by the probability distribution of X.

Formally,

$$\mathtt{E(Y) = \sum_{i} E(Y|X = x_i)Pr(X = x_i) = E[E(Y|X)]} \quad .$$

Law of iterated expectations

Prove the LIE for discrete random variables.

Law of iterated expectations

Remarks.

- This law implies that if E(Y|X) is equal to 0 then also E(Y) is equal to 0.
- This law applies also to expectations that are conditional on multiple random variables: E(Y)=E[E(Y|X,Z)], E(Y)=E[E(Y|X,Z,R)], ...

Statistical independence

When we deal with more than one rv an interesting question may rise: does knowing the value of some of them provide information about the others?

In our example: does knowing the weather conditions say something about the time Secchi will take to reach his office?

	Rain (X=0)	No rain (X=1)
Long commute (Y=50 min)	0.15	0.07
Short commute (Y=10 min)	0.15	0.63

Statistical independence

When we deal with more than one rv an interesting question may rise: does knowing the value of some of them provide information about the others?

In our example: does knowing the weather conditions say something about the time Secchi will take to reach his office?

	Rain (X=0)	No rain (X=1)
Long commute (Y=50 min)	0.15	0.07
Short commute (Y=10 min)	0.15	0.63

Statistical independence

Two random variables X and Y are independent if knowing the value of one of the two provides no information about the other.

Formally, X and Y are independently distributed if, for all values of x and y

$$\Pr(Y = y | X = x) = \Pr(Y = y) \quad .$$

Equivalently, since $\Pr(Y = y | X = x) = \frac{\Pr(X=x, Y=y)}{\Pr(X=x)}$

$$\Pr(Y = y, X = x) = \Pr(X = x) \Pr(Y = y) \quad ,$$

that is the joint distribution of two independent random variables is the product of their marginal distributions.

Statistical independence

Two random variables X and Y are independent if knowing the value of one of the two provides no information about the other.

Formally, X and Y are independently distributed if, for all values of x and y

$$\Pr(Y = y | X = x) = \Pr(Y = y) \quad .$$

Equivalently, since $\Pr(Y = y | X = x) = \frac{\Pr(X=x, Y=y)}{\Pr(X=x)}$

$$\Pr(Y = y, X = x) = \Pr(X = x) \Pr(Y = y) \quad ,$$

that is the joint distribution of two independent random variables is the product of their marginal distributions.

Statistical independence

Two random variables X and Y are independent if knowing the value of one of the two provides no information about the other.

Formally, X and Y are independently distributed if, for all values of x and y

$$\mathtt{Pr}(\mathtt{Y}=\mathtt{y}|\mathtt{X}=\mathtt{x})=\mathtt{Pr}(\mathtt{Y}=\mathtt{y})\ .$$

Equivalently, since $\mathtt{Pr}(\mathtt{Y}=\mathtt{y}|\mathtt{X}=\mathtt{x})=\frac{\mathtt{Pr}(\mathtt{X}=\mathtt{x},\mathtt{Y}=\mathtt{y})}{\mathtt{Pr}(\mathtt{X}=\mathtt{x})}$

$$\mathtt{Pr}(\mathtt{Y}=\mathtt{y},\mathtt{X}=\mathtt{x})=\mathtt{Pr}(\mathtt{X}=\mathtt{x})\,\mathtt{Pr}(\mathtt{Y}=\mathtt{y})\ ,$$

that is the joint distribution of two independent random variables is the product of their marginal distributions.

Covariance and correlation

How do we measure the extent to which two random variables move together?

The covariance between two random variable is defined as

$$
\begin{aligned}
&\mathrm{cov}(X, Y) = \sigma_{XY} = E\left[(X - \mu_X)(Y - \mu_Y)\right] = \\
&= \sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y)\Pr(X = x_i , Y = y_j) \ ,
\end{aligned}
$$

and it indeed measures the extent to which X and Y move together.

Covariance and correlation

How do we measure the extent to which two random variables move together?

The covariance between two random variable is defined as

$$\mathrm{cov}(X, Y) = \sigma_{XY} = E\left[(X - \mu_X)(Y - \mu_Y)\right] =$$
$$= \sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y)\Pr(X = x_i, Y = y_j) \ ,$$

and it indeed measures the extent to which X and Y move together.

Covariance and correlation

Interpretation. Suppose that when X is greater than its mean (X-μ_X>0) then also Y tends to be greater than its mean (Y-μ_Y>0) and when X is less than its mean (X-μ_X>0) also Y tends to be less than its mean(Y-μ_Y>0). In this case the products of (X-μ_X) and (Y-μ_Y) will be positive and the covariance as well.

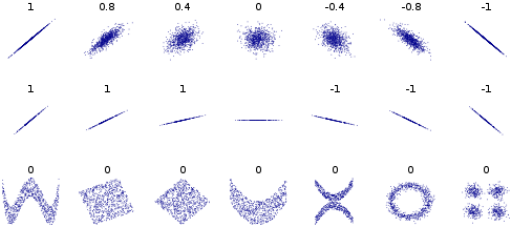
Covariance and correlation

To get rid of the problem concerning their units of measure we define the correlation as

$$\mathrm{corr}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} ,$$

which is a unitless coefficient. Note that,

- $-1 \leq \mathrm{corr}(X, Y) \leq 1$
- if X and Y are independent then corr(X,Y)=cov(X,y)=0, the converse is not necessarily true.
- corr(X,Y) captures only the linear dependence.



Sums of random variables

Some useful expressions (and exercises for you)

$$E(a + bX + cY) = a + b\mu_X + c\mu_Y \quad ,$$

$$\mathrm{var}(aX + bY) = a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2 \quad ,$$

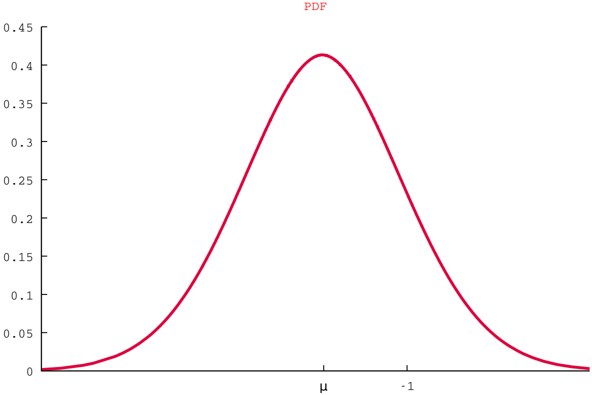
$$\mathrm{var}(Y_1 + Y_2 + \ldots + Y_n) = \sum_i \mathrm{var}(Y_i) + \sum_i \sum_{j \neq i} \mathrm{cov}(Y_i, Y_j) \quad ,$$

$$\mathrm{cov}(a + bX + cR, Y) = b\sigma_{XY} + c\sigma_{RY} \quad ,$$

$$|\sigma_{XY}| \leq \sigma_X \sigma_Y \quad .$$

Popular distributions in Econometrics

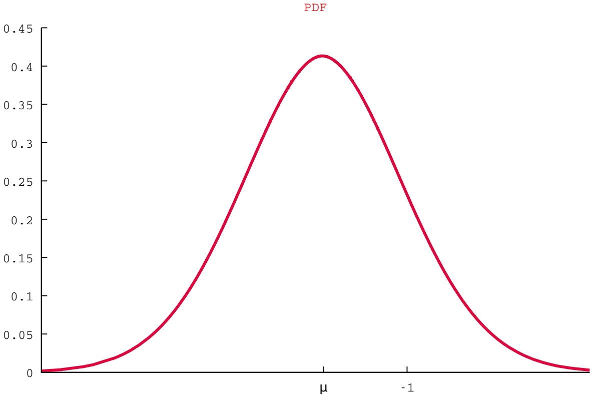
Normal distribution. The PDF of a normally distributed rv Y with mean μ and variance σ^2 looks like



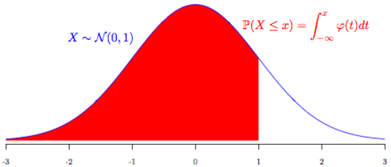
- What's the probability of observing a realization of Y lower than μ?
- What's the probability of observing a realization of Y lower than -1?

Popular distributions in Econometrics

Normal distribution. The PDF of a normally distributed rv Y with mean μ and variance σ^2 looks like



- What's the probability of observing a realization of Y lower than μ? Easy, the distribution is symmetric.
- What's the probability of observing a realization of Y lower than -1? Without a computer less easy but we have the statistical tables.



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441

This table reports values of the CDF of Z, a standardized normally distributed random variable. Z is obtained with $Z = \frac{Y - \mu}{\sigma}$ and by construction is distributed according to a Gaussian with 0 mean and variance 1.

Popular distributions in Econometrics

What's the probability of observing a realization of a Normally distributed rv Y (with mean μ and sd σ) lower than -1? To answer our question we need to follow a two-step procedure:

1. we need to obtain the standardized value of -1, $Z = \frac{(-1) - \mu}{\sigma}$
2. look for it in the table.

Ex. Assume $\mu = -2$ and $\sigma = 1$, then the standardized version of -1 reads $\frac{-1 - (-2)}{1} = 1$, so the probability of observing a realization of Y lower than 1 is 0.8413.

Popular distributions in Econometrics

What's the probability of observing a realization of a Normally distributed rv Y (with mean μ and sd σ) lower than -1? To answer our question we need to follow a two-step procedure:

1. we need to obtain the standardized value of -1, $Z = \dfrac{(-1) - \mu}{\sigma}$
2. look for it in the table.

Ex. Assume $\mu = -2$ and $\sigma = 1$, then the standardized version of -1 reads $\dfrac{-1 - (-2)}{1} = 1$, so the probability of observing a realization of Y lower than 1 is 0.8413.

Popular distributions in Econometrics

What's the probability of observing a realization of a Normally distributed rv Y (with mean μ and sd σ) lower than -1? To answer our question we need to follow a two-step procedure:

1. we need to obtain the standardized value of -1, $Z = \dfrac{(-1) - \mu}{\sigma}$
2. look for it in the table.

Ex. Assume $\mu = -2$ and $\sigma = 1$, then the standardized version of -1 reads $\dfrac{-1 - (-2)}{1} = 1$, so the probability of observing a realization of Y lower than 1 is 0.8413.

Popular distributions in Econometrics

What's the probability of observing a realization of a Normally distributed rv Y (with mean μ and sd σ) lower than -1? To answer our question we need to follow a two-step procedure:

1. we need to obtain the standardized value of -1, $Z = \dfrac{(-1) - \mu}{\sigma}$
2. look for it in the table.

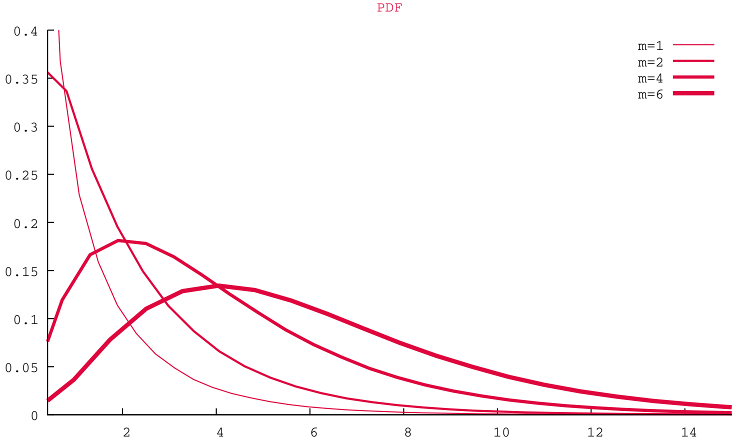
Ex. Assume $\mu = -2$ and $\sigma = 1$, then the standardized version of -1 reads $\dfrac{-1 - (-2)}{1} = 1$, so the probability of observing a realization of Y lower than 1 is 0.8413.

Popular distributions in Econometrics

Chi-squared. If Z_1, Z_2, Z_3,... are independent standard normal random variables then $Z_1^2 + Z_2^2 + Z_3^2 + \ldots$ is distributed according to a Chi-squared with m degrees of freedom, where m is the number of terms added together.

Popular distributions in Econometrics

Chi-squared. If $Z_1, Z_2, Z_3, \ldots$ are independent standard normal random variables then $Z_1^2 + Z_2^2 + Z_3^2 + \ldots$ is distributed according to a Chi-squared with m degrees of freedom, where m is the number of terms added together.

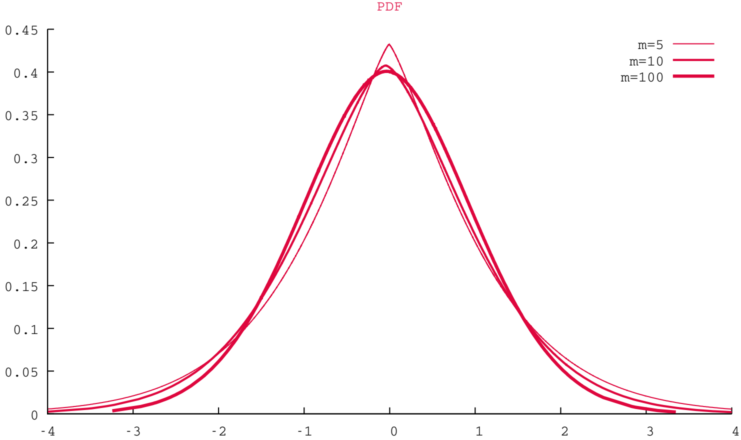


Popular distributions in Econometrics

Student's t. If Z is a standard normal random variables and W an independent Chi-squared random variable with m degrees of freedom then $\frac{Z}{\sqrt{W/m}}$ is distributed according to a Student t distribution.

Popular distributions in Econometrics

Student's t. If Z is a standard normal random variables and W an independent Chi-squared random variable with m degrees of freedom then $\frac{Z}{\sqrt{W/m}}$ is distributed according to a Student t distribution.



Almost all the econometric procedures we will be using involve averages or weighted averages.

Almost all the econometric procedures we will be using involve averages or weighted averages.

Using the textbook wording we distinguish between

- long-run: weights are the true probabilities of each realization;
- short-run: weights are the observed probabilities of each realization.

Almost all the econometric procedures we will be using involve averages or weighted averages.

Using the textbook wording we distinguish between

- long-run: weights are the true probabilities of each realization; [population mean]
- short-run: weights are the observed probabilities of each realization. [sample average]

Almost all the econometric procedures we will be using involve averages or weighted averages.

Using the textbook wording we distinguish between

- long-run: weights are the true probabilities of each realization; [population mean]
- short-run: weights are the observed probabilities of each realization. [sample average]

Characterizing the distributions of sample averages is an essential step to assess the performance of each econometric procedures.

Section 3

Random sampling

Random sampling

In a simple random sample, n objects are drawn at random from a population where

- each object has the same probability of being drawn;
- the value of the random variable Y for the i_{th} randomly drawn object is denoted Y_i;
- since each object is equally likely to be drawn and the distribution of Y_i is the same for for all i the random variables $Y_1,\dots,Y_n$ are said to be independently and identically distributed (i.i.d.).

Example. Imagine an urn with 3 balls, one white, one red and one blue. List different valid random samples with n=1, n=2 and n=3.

Random sampling

In a simple random sample, n objects are drawn at random from a population where

- each object has the same probability of being drawn;
- the value of the random variable Y for the i_{th} randomly drawn object is denoted Y_i;
- since each object is equally likely to be drawn and the distribution of Y_i is the same for for all i the random variables $Y_1,\ldots,Y_n$ are said to be independently and identically distributed (i.i.d.).

Example. Imagine an urn with 3 balls, one white, one red and one blue. List different valid random samples with n=1, n=2 and n=3.

Distribution of the sample average

The sample average, $\overline{Y}$ of the n obs $Y_1,\ldots,Y_n$ is

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad ,$$

where because $Y_1,\ldots,Y_n$ are random then also their average is random. The value of $\overline{Y}$ differs from one randomly drawn sample to the other. It is itself a rv.

Since $\overline{Y}$ is a random variable, it has a probability distribution, called sampling distribution because it is the probability distribution associated with the different values of $\overline{Y}$ obtained from different random samples.

Distribution of the sample average

The sample average, $\overline{Y}$ of the n obs $Y_1,\dots,Y_n$ is

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad ,$$

where because $Y_1,\dots,Y_n$ are random then also their average is random. The value of $\overline{Y}$ differs from one randomly drawn sample to the other. It is itself a rv.

Since $\overline{Y}$ is a random variable, it has a probability distribution, called sampling distribution because it is the probability distribution associated with the different values of $\overline{Y}$ obtained from different random samples.

Distribution of the sample average

We start by characterizing the mean and the variance of $\overline{Y}$. Let $Y_1,\ldots,Y_n$ be i.i.d and let μ_Y and σ_Y^2 denote the mean and the variance common to any Y_i. Then

$$
\begin{aligned}
E(\overline{Y}) &= E\left(\frac{1}{n}\sum_i Y_i\right) = \frac{1}{n}\sum_i E(Y_i) = \mu_Y \\
var(\overline{Y}) &= var\left(\frac{1}{n}\sum_i Y_i\right) = \frac{1}{n^2}\sum_i var(Y_i) = \frac{\sigma_Y^2}{n} \\
std.dev(\overline{Y}) &= \frac{\sigma_Y}{\sqrt{n}} \quad .
\end{aligned}
$$

Remarks. This result hold whatever the distribution of Y_i, that is we do not need to assume that Y_i is normally distributed.

Large samples

There are two ways to characterize the distribution of the sample average:

- exact: one assumes the distribution of Y and derives the distribution of $\bar{Y}$ for any n;
- approximate: one assumes n (sample size) to be large and approximates the distribution of $\bar{Y}$ using statistical tools.

We now discuss two key tools used to approximate sampling distribution when the sample size is large: the Law of Large Numbers and the Central Limit Theorem.

Large samples

There are two ways to characterize the distribution of the sample average:

- exact: one assumes the distribution of Y and derives the distribution of $\bar{Y}$ for any n;
- approximate: one assumes n (sample size) to be large and approximates the distribution of $\bar{Y}$ using statistical tools.

We now discuss two key tools used to approximate sampling distribution when the sample size is large: the Law of Large Numbers and the Central Limit Theorem.

The Law of Large Numbers (LLN)

To get the intuition of the LLN let us consider a rv "Commuting time" with its distribution:

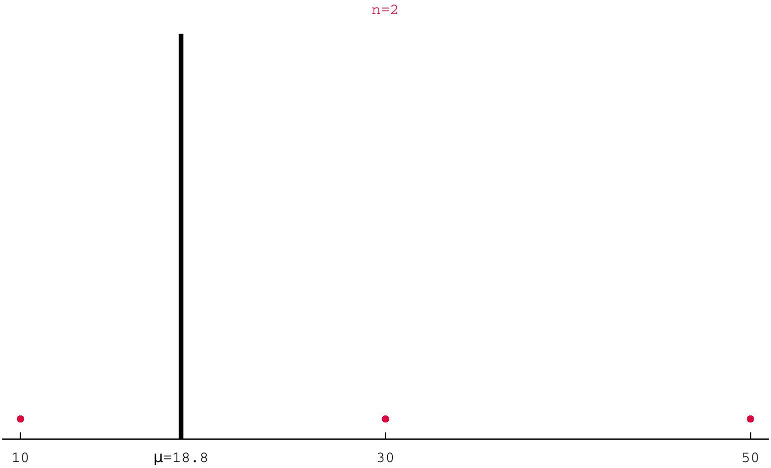
Table: Marginal distribution of Commuting times

	P(Y)
Long commute (Y=50 min)	0.22
Short commute (Y=10 min)	0.78

Draw 1000 random samples with n=2 from the distribution of Y:

- how many values can $\overline{Y}$ can take on?
- Is either of them close to the true $\mu_Y = 18.8$?

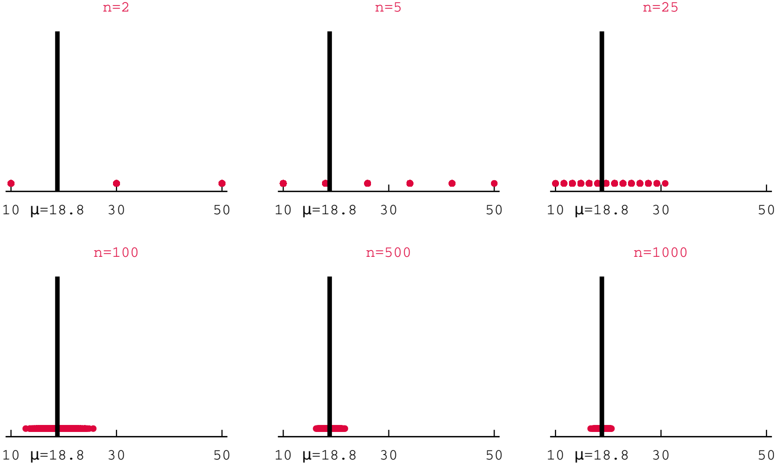
The Law of Large Numbers (LLN)



The Law of Large Numbers (LLN)

The LLN tells you what happens if you increase n, the sample size. It states that, under general conditions, $\bar{Y}$ will be near to μ_Y with very high probability when n is large.

The Law of Large Numbers (LLN)



The Law of Large Numbers (LLN)

LLN. The LLN says that if Y_i ($i = 1, \ldots, n$) are i.i.d. with $E(Y)_i = \mu_Y$ and if extreme values are unlikely (technically if $var(Y_i) < \infty$) then $\bar{Y}$ converges in probability to μ_Y, that is

$$\bar{Y} \xrightarrow{p} \mu_Y \quad .$$

This means that the probability that $\bar{Y}$ is in the range $(\mu_Y - c, \mu_Y + c)$ becomes arbitrarily close to 1 as n increases for any c>0.

The Central Limit Theorem (CLT)

Let's go back to our example with n=2.

- Can we say something more about the whole shape of the sampling distribution of $\overline{Y}$?

- Does a standardized Normal distribution (with mean 0 and sd 1) fit well the sampling distribution of $\overline{Y}$, once we normalize it as

$$\overline{Z} = \frac{\overline{Y} - \mu_Y}{\sqrt{\frac{\sigma_Y^2}{n}}}$$

with $n = 2$, mean $\mu_Y = 18.8$ and variance $\sigma_Y^2/n = 0.0858$?

The Central Limit Theorem (CLT)

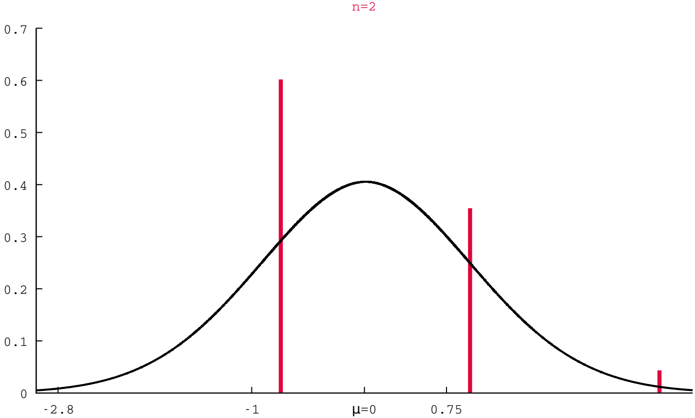
Let's go back to our example with n=2.

- Can we say something more about the whole shape of the sampling distribution of $\overline{Y}$?
- Does a standardized Normal distribution (with mean 0 and sd 1) fit well the sampling distribution of $\overline{Y}$, once we normalize it as

$$\overline{Z} = \frac{\overline{Y} - \mu_Y}{\sqrt{\frac{\sigma_Y^2}{n}}}$$

with $n = 2$, mean $\mu_Y = 18.8$ and variance $\sigma_Y^2/n = 0.0858$?

The Central Limit Theorem (CLT)



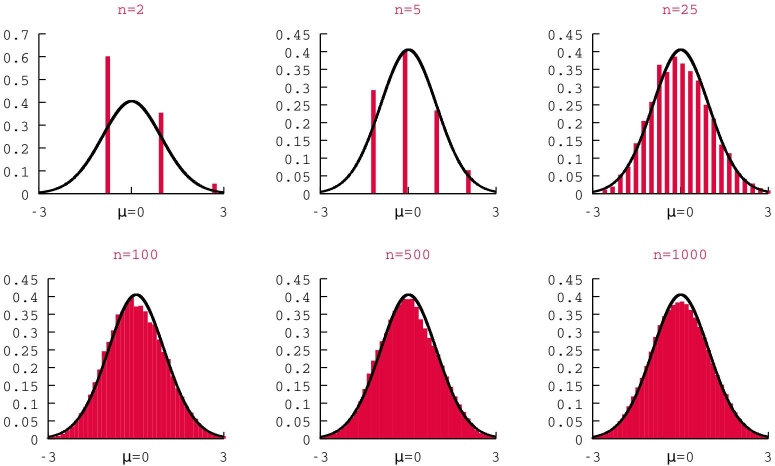
The Central Limit Theorem (CLT)

The CLT states that, under general conditions, the sampling distribution of $\overline{\mathtt{Y}}$ is well approximated by a Normal distribution with mean $\mu_{\mathtt{Y}}$ and variance $\sigma_{\mathtt{Y}}^2/\mathtt{n}$ when n is large.

The Central Limit Theorem (CLT)

In other words the CLT states that, under general conditions, the sampling distribution of $\bar{Z} = \frac{\bar{Y} - \mu_Y}{\sqrt{\frac{\sigma_Y^2}{n}}}$ is well approximated by a standardized Normal distribution when n is large.

The Central Limit Theorem (CLT)



The Central Limit Theorem (CLT)

CLT. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. with $E(Y_i) = \mu_Y$ and $var(Y_i) = \sigma_Y^2 < \infty$.As $n \to \infty$ the distribution of

$$\frac{\overline{Y} - \mu_Y}{\frac{\sigma_Y}{\sqrt{n}}}$$

becomes arbitrarily well approximated by the standard normal distribution. We write

$$\frac{\overline{Y} - \mu_Y}{\frac{\sigma_Y}{\sqrt{n}}} \xrightarrow{d} N(0,1) \quad .$$

Practice

Exercise 5. Let X be a Bernoulli rv with Pr(X=1)=0.9, Y$\sim$N(0,4) and W$\sim$N(0,16). Assume that X, Y and W are independent and let S=XY+(1-X)W. Show that:

- E[Y^2]=4 and E[W^2]=16;
- E[Y^3]=0 and E[W^3]=0;

and compute

- E[S], E[S^2] and E[S^3].

  [hint: use the law of iterated expectations]

Practice

Exercise 6. Suppose Y_i, with $i = 1, \ldots, n$, are independent rv each distributed as N(10,4). Using the CLT compute $\Pr(9.6 \leq \bar{Y} \leq 10.4)$ when n=20, n=100 and n=1000.

Exercise 7. Let Y_i, with $i = 1, \ldots, n$, be Bernoulli rv with $\Pr(Y_i=1)=0.6$. compute

- $\Pr(\bar{Y}>0.64)$ when n=50;
- the sample size n such that $\Pr(0.56 \leq \bar{Y} \leq 0.64)=0.95$.

Section 4

PDF estimate [later in the semester]

Density estimation

As discussed so far the PDF is a fundamental concept in statistics. If you have a continous random variable X with a known PDF p(x) you can build probabilities associated with X using

$$\mathrm{Pr}(a \leq X \leq b) = \int_a^b dx\, p(x) \quad , \forall\, a < b \quad .$$

Now, suppose you do not know the true p(x) but that you have a set of observed data points assumed to be a sample from the unknown probability density function p(x).

Is there ways to recover the form of p(x)? This is what we try to do in the next few slides.

Density estimation

- Our aim is not to derive the full theory of density estimation but only to get the main intuition behind it to be able to correctly understand its strengths but also its limitations.

- Understanding the intuition is a necessary condition also to apply correctly the methods in real empirical investigations.

- Never use a method, technique or statistical tool without knowing how it works. It is always a source of problems!

Density estimation

There are two fundamental approaches to density estimation

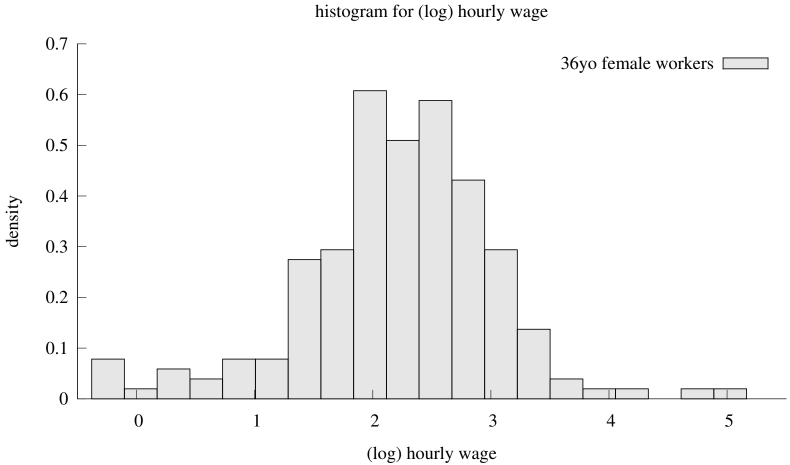
- Parametric. One assumes that the data are drawn from a known parametric family of distributions and then he estimates the corresponding parameters. ▸ Example.

- Non Parametric. One assumes only that the distribution has a probability density p(x). Then data will be allowed to speak for themselves in determining the shape and the properties of p(x). This is the approach we follow.

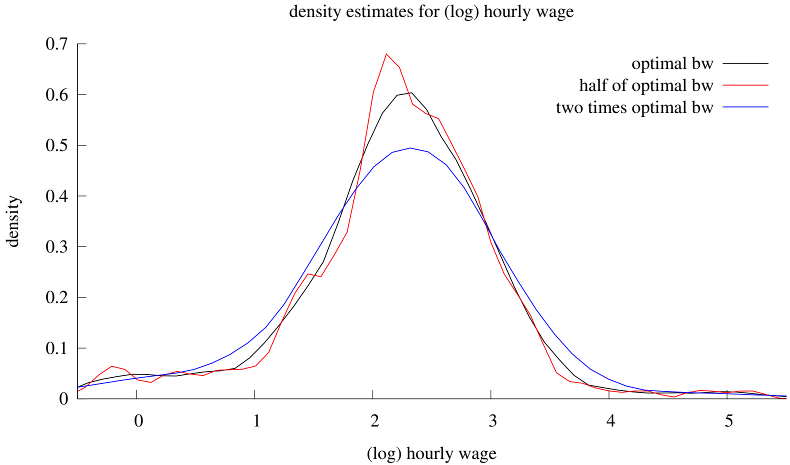
Density estimation

- A very natural use of density estimates is in the informal investigation of the properties of a given set of data.
- This kind of descriptive data investigation should forego any regression analysis.
- Let's consider an example.

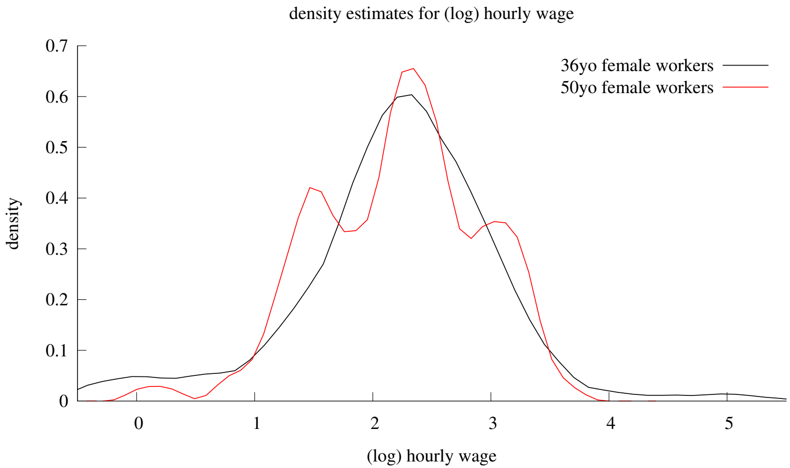
Hourly wage of a sample of American workers in 1993.



Hourly wage of a sample of American workers in 1993.



What can we do with this kernel density estimate?



In this example, but this is very often the case, the conclusions could only be regarded as a clue for further investigations.

Density estimation

The oldest and most used density estimator is the histogram, an estimate of the density formed by splitting the range of a variable X into equally spaced intervals and calculating the fraction of the sample falling in each interval.

Density estimation

Practically to build an histogram one has to set:

1. origin: x_0
2. bin width: h
3. bins: defined as $[x_0 + m(2h), x_0 + (2h)(m+1))$ where $m = \{\text{positive and negative integers}\}$.

Given a sample $\{x_i, i = 1, \ldots, n\}$ the histogram $\hat{f}(x)$ is then defined as

$$\hat{p}(x) = \frac{1}{2h}\left(\frac{\#\text{ of } x_i \text{ in the same bin as } x}{n}\right) \quad ,$$

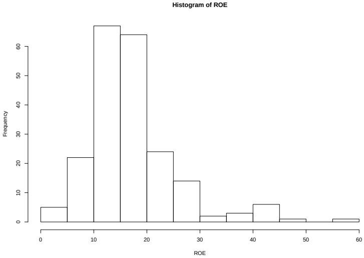
or, if one assumes a varying bin width,

$$\hat{p}(x) = \frac{1}{\text{width of the bin}}\left(\frac{\#\text{ of } x_i \text{ in the same bin as } x}{n}\right) \quad .$$

R practitioner corner

In R it is very easy to obtain an histogram of a variable.

```
> library(foreign)
> ceosal1<-read.dta("
http://fmwww.bc.edu/ec-p/data/wooldridge/ceosal1.dta
>  ")
> ROE <- ceosal1$roe
> hist(ROE,xlim=c(0,60))
```



Density estimation - Histograms

Why it is ever necessary to use methods more sophisticated than the simple histogram?

- general reason: very often you build density estimates as intermediate components of other methods, in these cases the discontinuity of the histograms is a problem

- specific reason: the shape of the histogram might strongly depends on the origin x_0. Example.

density estimation - naive estimator

Formally, consider estimation of the density p(x) of a scalar random variable X evaluated at x. Since the density is the derivative of the CDF P(x) then

$$p(x) = \lim_{h \to 0} \frac{P(x+h) - P(x-h)}{2h} = \lim_{h \to 0} \frac{\Pr[x-h < X < x+h]}{2h} \quad .$$

So for a sample $\{x_i, i = 1, \ldots, n\}$ a natural estimator for p(x) is

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1(x - h < x_i < x + h)}{2h} \quad ,$$

where $1(\cdot)$ is the indicator function.

density estimation - naive estimator

A way to express the estimator without using the indicator function is obtained by considering a weight function w

$$w(z) = \begin{cases} \frac{1}{2} & \text{if } |z| < 1 \\ 0 & \text{otherwise} \end{cases} .$$

Howto interpret w(z).

density estimation - naive estimator

Howto interpret w(z).

- If one evaluates w at the point $0 - x_i$:

$$w(0 - x_i) = \begin{cases} \frac{1}{2} & \text{if } |0 - x_i| < 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$$|0 - x_i| < 1 \Leftrightarrow \begin{cases} 0 - x_i < 1 \\ 0 - x_i > -1 \end{cases} \Leftrightarrow 0 - 1 < x_i < 0 + 1$$

which means that $w(0 - x_i)$ assigns a weight 1/2 to each x_i whose distance from 0 is less than 1.

density estimation - naive estimator

Howto interpret w(z).

- Evaluate w at the point $(x - x_i)/h$

$$w\left(\frac{x - x_i}{h}\right) = \begin{cases} \frac{1}{2} & \text{if } \left|\frac{x-x_i}{h}\right| < 1 \\ 0 & \text{otherwise} \end{cases}$$

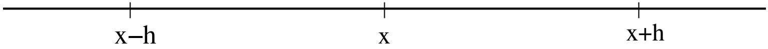
where

$$\left|\frac{x - x_i}{h}\right| < 1 \Leftrightarrow \begin{cases} x - x_i < h \\ x - x_i > -h \end{cases} \Leftrightarrow x - h < x_i < x + h$$

which means that $w(\frac{x-x_i}{h})$ assigns a weight 1/2 to each x_i whose distance from x is less than h.

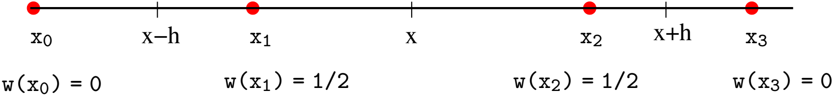
density estimation - naive estimator

`Howto interpret w(z).`



density estimation - naive estimator

Howto interpret w(z).



density estimation - naive estimator

With the function w(x) we define the naive estimator as

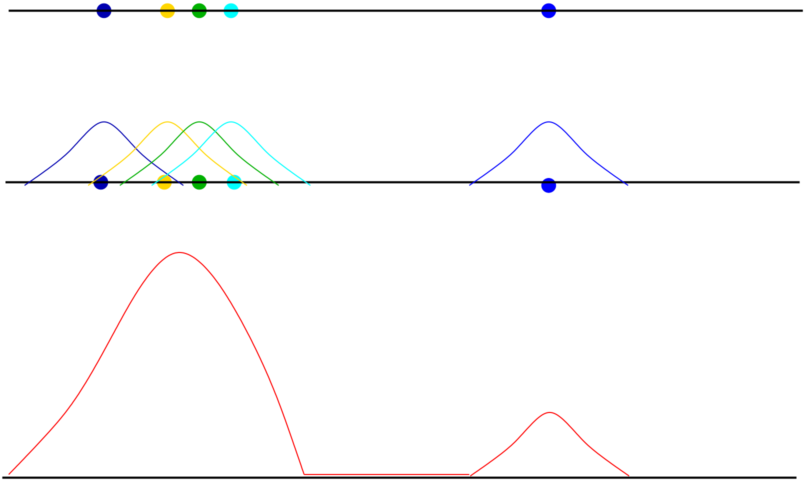
$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - x_i}{h}\right) \quad .$$

Intuition. The naive estimates consists in placing a "box" of width 2h and height $\frac{1}{(2nh)}$ on each observation and then summing up. ▸ Example.

Limitations. Its main limitation derives from the fact that it is not a continuous function having jumps at the points $x_i \pm h$.

density estimation - kernel estimator

A straightforward generalization of the naive estimator is obtained replacing the boxes with a more general function, called ▸ Kernel



density estimation - kernel estimator

The kernel density estimator is then defined as

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right) \quad ,$$

where h in this context is called bandwidth (or smoothing parameter) and K is any kernel function.

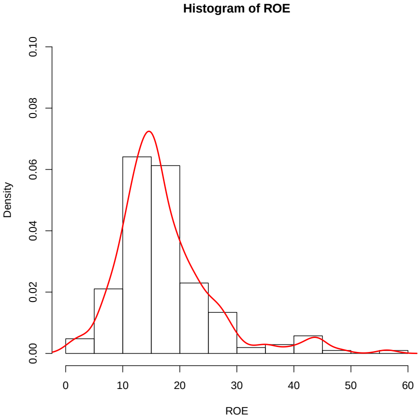
Most used kernel functions are symmetric univariate densities like Gaussian, Rectangular, etc. ▸ Kernel functions

Intuition: the kernel estimator is a sum of "bumps" placed at the observations. The kernel function K determines the shape of the bumps while the window width h determines their width.

R practitioner corner

In R it is very easy to obtain also the density estimate.

```
> library(foreign)
> ceosal1<-read.dta("
http://fmwww.bc.edu/ec-p/data/wooldridge/ceosal1.dta
>  ")
> ROE <- ceosal1$roe
> hist(ROE,xlim=c(0,60),freq=FALSE,ylim=c(0,0.10))
> lines(density(ROE,bw="nrd0",adjust=c(1),kernel=c('gaussian')),lwd=2,col="red")
```



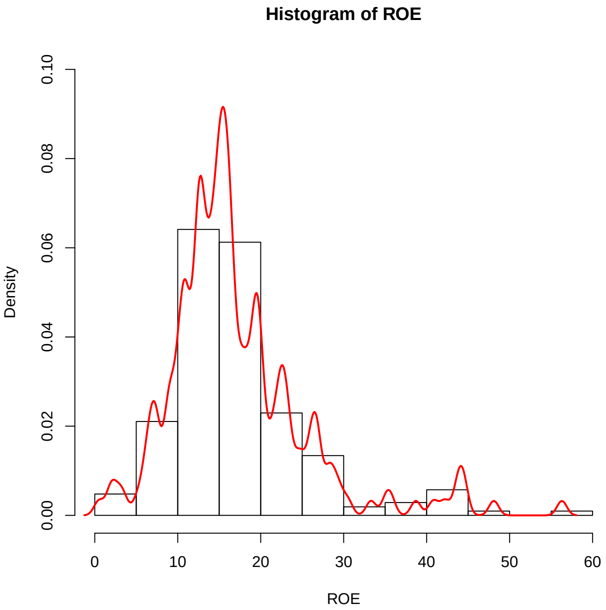
How much smoothing?

The problem of choosing how much to smooth (that is h) is of crucial importance in density estimation. This choice will always be influenced by the purpose for which the density estimate is to be used.

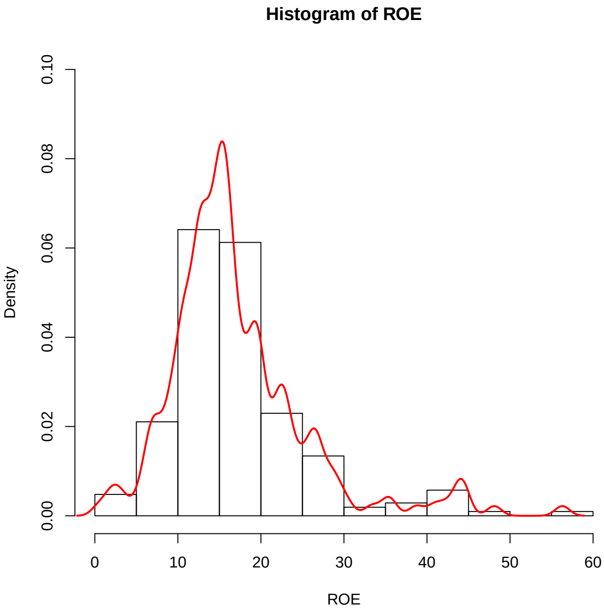
- representing data: subjective choice of the smoothing parameter

- presenting conclusions: automatic bandwidth selection, in any case it is better to undersmooth (smoothing by eye is easier than unsmooth)

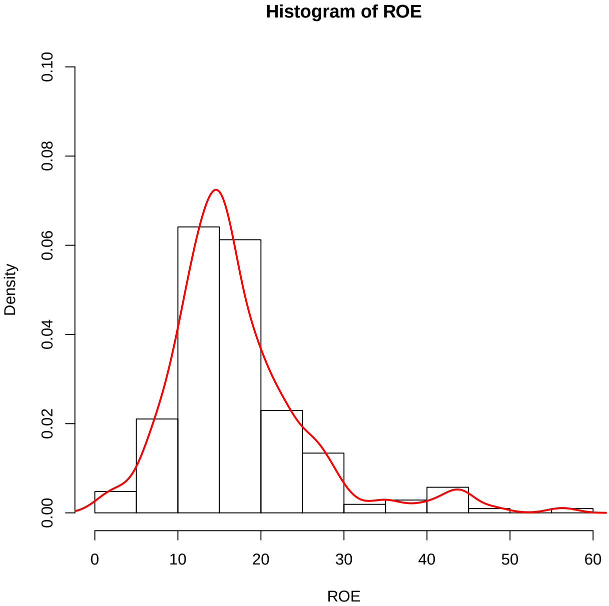
how much smoothing?



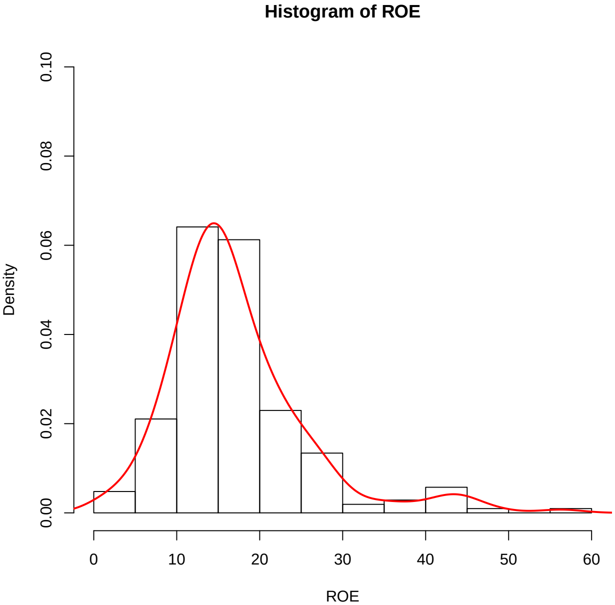
how much smoothing?



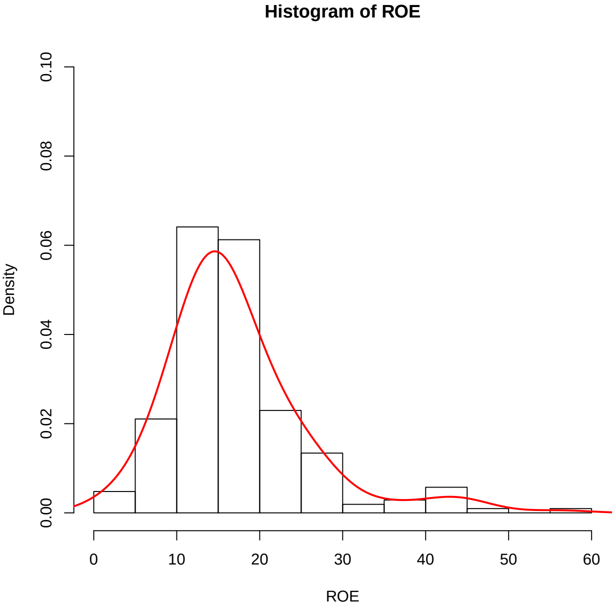
how much smoothing?



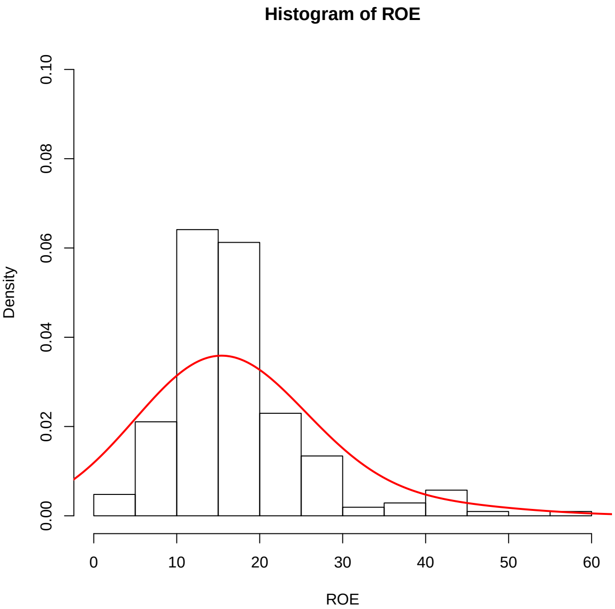
how much smoothing?



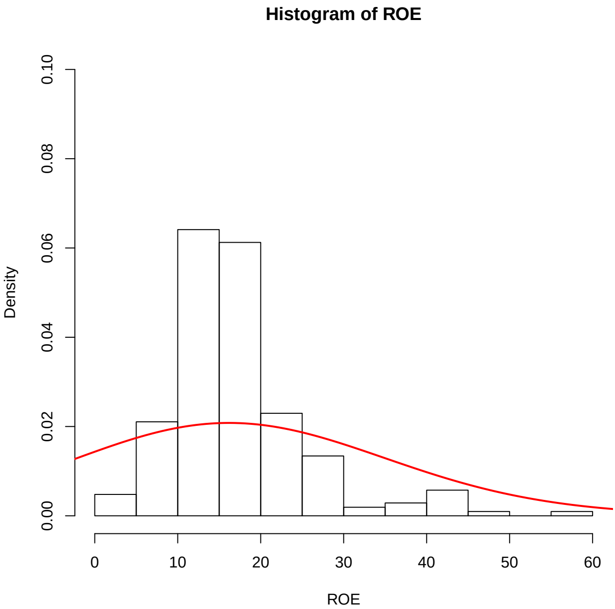
how much smoothing?



how much smoothing?

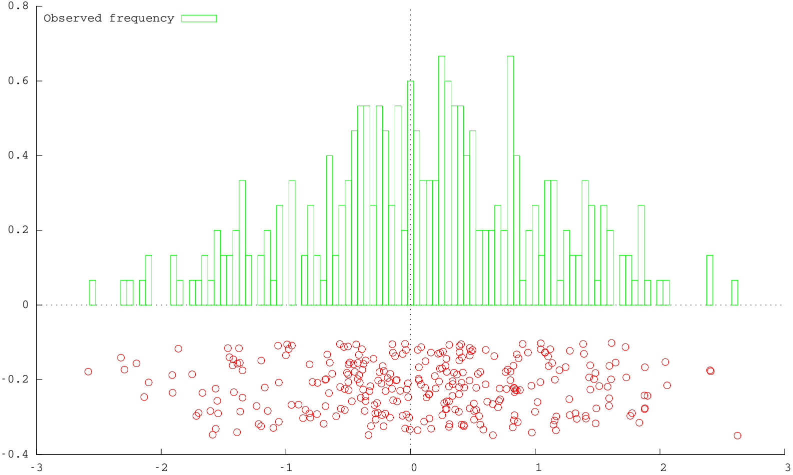


how much smoothing?



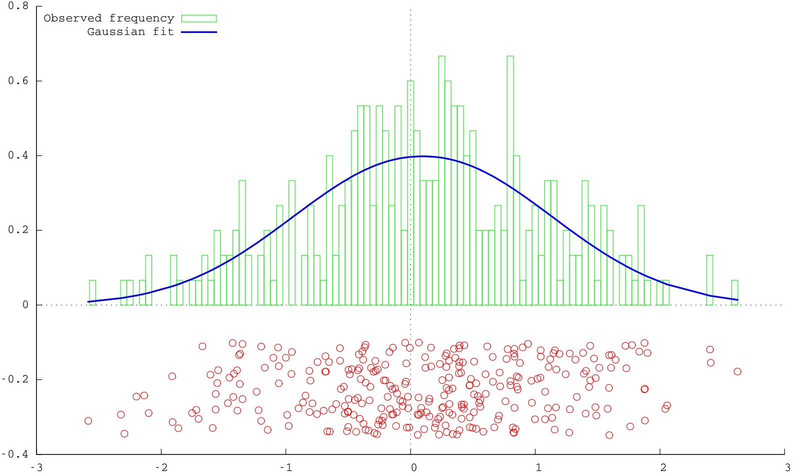
Hyper-references

Parametric DE



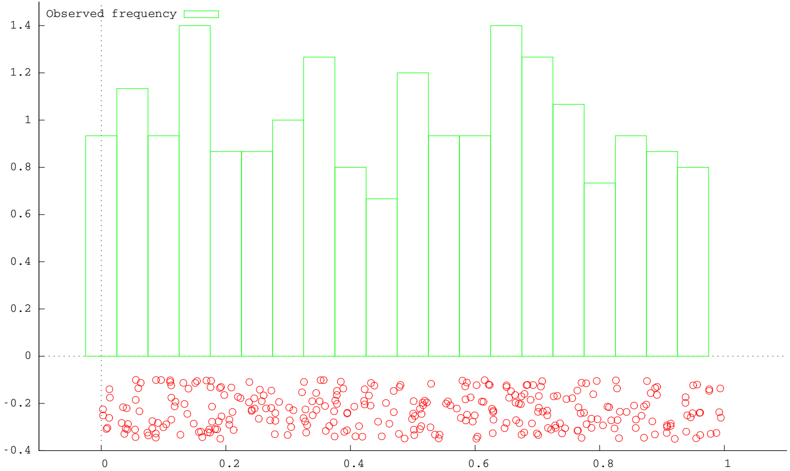
Back

Parametric DE



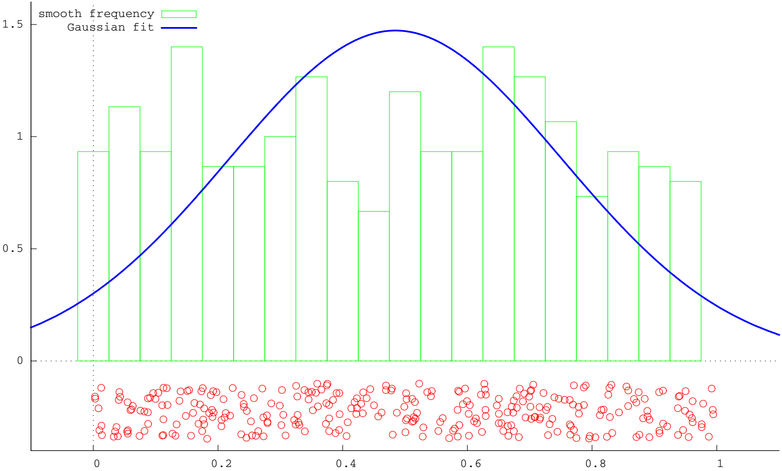
Back

Parametric DE



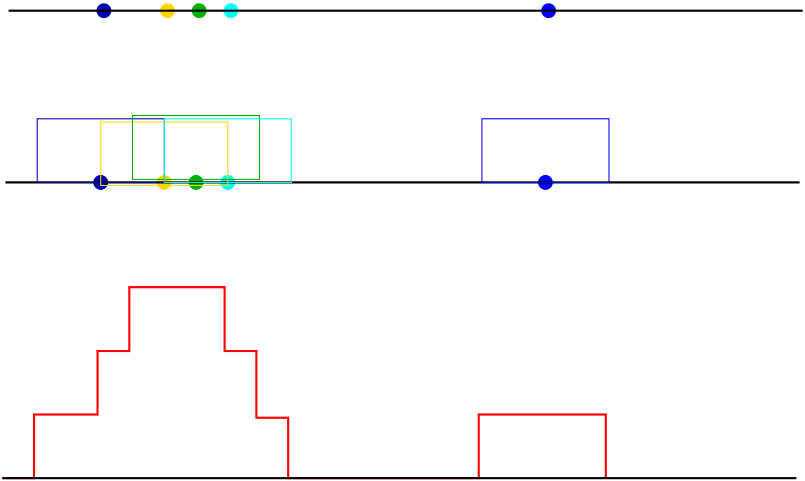
Back

Parametric DE



Back

Naive estimator



Back

kernel functions

A kernel is a non-negative real-valued integrable function K satisfying the following two requirements:

$$\int_{-\infty}^{+\infty} K(u)\,du = 1$$

$K(-u) = K(u)$ for all values of u .

These two requirements ensure that

- a density estimate based on K(s) will be a PDF;
- the average of the estimated density is equal to the sample average.

▸ Back