

Inferential Statistics

L0 - Introduction and probability theory

Erlis Ruli (erlis.ruli@unipd.it)

Department of Statistics, University of Padova

Contents

- 1 Course Info
- 2 What is Statistics ?
- 3 Three statistical problems
- 4 A quick dive into Probability

Info

Me: Erlis Ruli (erlis.ruli@unipd.it), Department of Statistics.

The course: 6 credits, 48 hours, 24 lectures overall.

Prerequisites: Probability Theory, Mathematical Analysis, Linear Algebra.

Time table: Mon 14:30-16:30 in room Ce and Fri 08:30-10:30 in room Ae; more details on Moodle.

Each lecture: covers theory + problems. For some problems we'll use electronic calculation (\mathbb{R}^1) in BYOD.

Learning disabilities: I'll be happy to help you, just let me know.

¹<https://cran.r-project.org/>

Info

Exam sessions: 25/01/2023, 14/02/2023, jun-2024, jul-2024, sep-2024.

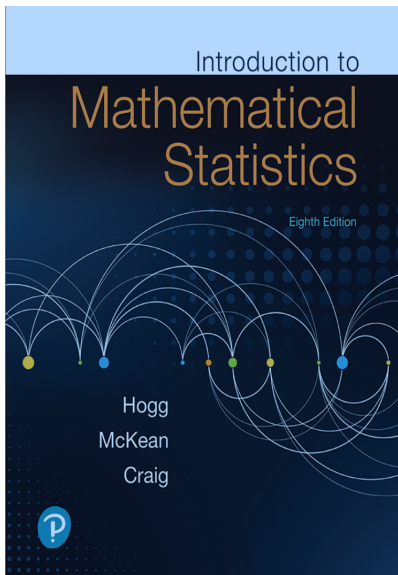
Exam: Written closed-book exam, 2-4 questions with subquestions of varying difficulty; you can bring your PC/hand calculator and a single page (A4 format) with your notes, formulas, etc.

Performance evaluation: You must show that you master the concepts and you are able to apply the methods to simple problems, terminology, etc.

Grade: Exam performance + extra points based on classroom engagement.
Classroom engagement evaluation criteria:

- based on the degree of your **active** classroom participation, (i.e. answer/ask questions, solve homework exercises);
- overall point a scalar in $[0, 3]$;
- will be assigned near the end of the course and valid for the A.Y.

Textbook and other materials



Other materials:

- Kupper et al. (2003) Exercises and Solutions in Statistical Theory, CRC Press, has a lot of interesting problems.
- handouts, slides, solved problems, past exams, on moodle.

Other interesting books:

- Wasserman (2003) All of Statistics, Springer.
- Casella and Berger (2002) Statistical Inference, 2nd ed, Duxbury.

Course structure

- L0: Background on probability theory (part I)
- L1: Background on probability theory (part II)
- L2: Statistical models
- L3: Likelihood function
- L4: Point estimation
- L5: Hypothesis testing
- L6: Confidence sets
- L7: Bayesian statistics (if time permits)
- L8: Nonparametric statistics (if time permits)

Some lectures will make a heavy use of R; if possible take your laptop always with you!

Why Statistics?

Science is essentially **experimental**². Indeed, scientific knowledge can only be "proved" through experiments and data.

In Science, we often wish to provide:

- (i) evidence about (or against) some theory
- (ii) support for a business decision problem.

An instance of (ii): you may produce³ washing machines (WMs) and you wish to measure the amount of energy consumption.

²Leave Math alone.

³Essentially all major household appliances sold in the EU market must be accompanied by a technical documentation describing: energy consumption, water consumption, noise emission, etc.; shorturl.at/PR015

A common approach

Data are collected through experiments run in a lab, following a specific protocol.

Suppose there are n measurements of energy consumption. Typically the values are average out to get a number, say 1.1 kw/h.

A typical question then is: How precise is this value?

Not an obvious answer and often an interval of the type

$$1.1 \pm \text{tolerance}$$

is computed.

In this course we'll learn what justifies this rule and many other alternative approaches.

Inferential Statistics ↔ Decision: the roadmap

In a nutshell, Inferential Statistics involves three steps:

- (i) translate the problem (verification of theory, decision problem, etc.) in terms of a **statistical model**
- (ii) fit the statistical model to the data
- (iii) translate results back in terms of the original problem: iterate through (i)-(iii) if necessary.

There are two key points here:

- (a) choice the statistical model
- (b) fit the statistical model to the data.

"All models are wrong, but some are useful" (G. Box)

Realistically, the true model is **unknown** and, it may or may not coincide with our choice.

However, carefully chosen models typically provide good and useful approximations.⁴

In this course we'll have a quick look at how to build simple models (a) but we'll go deeper in the **methods** part (b).

In particular, given a statistical model, we'll see:

- methods for fitting a model to the data (model estimation)
- methods and principles for getting evidence from data (inference or point (iii) above).

Here are three typical practical problems...

⁴Models as approximations of reality are ubiquitous in science, e.g. Rutherford (Physics), Mendel (Genetics), Weber (Sociology), Smith (Economics), etc.

Example 1: Estimation

Every WM sold in the UE market must be accompanied by a document that describes, among other things, the energy consumed during a typical washing cycle.

To measure energy consumption, the WM is tested in laboratory several times, following a precise protocol. With the observed values of energy consumption the manufacturer wishes to estimate or learn the most “plausible value”.

This is an estimation problem: to be addressed in L. 4.

Example 2: Hypothesis testing

It is claimed that to reduce energy consumption, current WM motors must be replaced by a newer model. WMs with the current motor and with the new motor are tested and their energy consumption is measured. Based on the observed data we must tell if the claim is true or not.

This is an hypothesis testing problem: to be addressed in L. 5.

Example 3: Confidence set

The manufacturer must also declare a measure of variability for this estimate, i.e. it has to declare tolerance limits within which the consumed energy is expected to vary with high probability.

This is a confidence set problem: to be addressed in L. 6.

Many views, so many methods!

Statistical inference can be cross-classified as

Paradigm	Setting	
	Parametric	Nonparametric
Frequentist	(i)	(ii)
Bayesian	(iii)	(iv)

(i) and (iii) typically use probability models with a **finite number** of parameters. The nonparametric setting can be seen as founded on probability models with infinitely many parameters.

In this course we'll focus mainly on (i) and if time permits, we'll see a bite of (iii) and (ii).

Probability

In Probability Theory we answer questions s.a.

- what is the pr. that the sum of two dice equals 4?
- what is the pr. to see an electron in a certain region of the electron shell?
- ...

To answer these questions we need to do:

- (1) set up a **probability model**, e.g. fix a sample space, a probability measure and
- (2) follow the **rules of probability**.

We are going to touch on each of these below.

The probability model

Choosing a probability model means defining the random experiment and the possible outcomes.

For instance, when throwing two dice, we may assume that each die:

- is balanced and
- always show face up.

Sample space

Sample space \mathcal{S} : space of all realisations s of a rnd experiment.

\mathcal{S} may be: (i) discrete (finite or infinite), (ii) continuous (infinite).

Examples of discrete \mathcal{S} :

- # faces of die (finite)
- # defective items in a lot of 100 items (finite)
- # tosses of a coin until the first head appears (infinite)

Examples of continuous \mathcal{S} :

- power of a circuit at time t
- position of an electron of a hydrogen atom at time t
- the price of BTC tomorrow at 08:53, etc.

Events

Sample point s : an element of \mathcal{S}

Event E : a subset of \mathcal{S} .

Example 1

In the two-dice experiment

$$\mathcal{S} = \{(1, 1), (1, 2), \dots, (6, 6)\},$$

and $s = (1, 1)$ is the sample point “both dies show face 1”.

The event “both numbers are less or equal to two” is

$$\mathcal{S} = \{(1, 1), (1, 2), (2, 1), (2, 2)\}.$$

There are two **special events**: \mathcal{S} and $\emptyset = \mathcal{S}^c$.

Take all possible events and put them in a safe place, say \mathcal{A} .

Example 2

For $\mathcal{S} = \{0, 1, 2\}$, possible events are $\{0\}, \{0, 1\}, \dots$; all in a single set^a

$$\mathcal{A} = \{\emptyset, \{0\}, \{1\}, \{2\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, \{0, 1, 2\}\}.$$

^aYes, $\mathcal{A} = \mathcal{P}(\mathcal{S})$.

Summing up, if E is any event:

- $E \subset \mathcal{S}$ but
- $E \in \mathcal{A}$.

In continuous \mathcal{S} we have to be more careful at defining \mathcal{A} , it may contain "undesirable" subsets.⁵

⁵We say that an event E is a **measurable** subset of \mathcal{S} , where "measurable" means a set that belongs to a suitable sigma field of \mathcal{S} .

$$P(E)$$

Probability is a function $P : \mathcal{A} \rightarrow [0, 1]$ s.t.

- (i) $P(\mathcal{S}) = 1$
- (ii) $P(E) \geq 0$ for every event E
- (iii) If E_1, E_2, \dots , s.t. $E_i \cap E_j = \emptyset$ for all i, j then $P(\cup E_i) = \sum_i P(E_i)$.

Example 3 (Single die problem)

The probability that it will face up 0 is zero. Indeed, if we let $E =$ "the die shows face 0", then $E = \emptyset$ and

$$1 = P(\mathcal{S}) = P(\mathcal{S} \cup \emptyset) = P(\mathcal{S}) + P(\emptyset),$$

thus $P(\emptyset) = 0$.

Example 4 (Two dice problem)

Compute the probability that at least one face shows up 1. Let

E = "at least one face shows up 1"

E_1 = "only first die shows up 1"

E_2 = "only second die shows up 1"

E_3 = "both dies show up 1".

We see that $E = E_1 \cup E_2 \cup E_3$ and $E_i \cap E_j = \emptyset$, thus

$$P(E) = P(E_1) + P(E_2) + P(E_3) = 5/36 + 5/36 + 1/36 = 11/36.$$

We could have computed this probability by counting all the simple events in E .

Conditional probability and independence

Given two events E, A :

Conditional probability of the E given A :

$$P(E|A) = P(E \cap A)/P(A),$$

provided $P(A) > 0$.

E is independent of A iff $P(E|A) = P(E)$.

For more than two events, say E, A, B :

- if $P(E|A) = P(E)$ we can **only** say A is independent of E
- to have **complete independence** of E, A, B we need to have

$$P(E \cap A) = P(E)P(A), \quad P(A \cap B) = P(A)P(B),$$

$$P(E \cap B) = P(E)P(B)$$

$$P(E \cap A \cap B) = P(E)P(A)P(B).$$

Random variables

The triple $(\mathcal{S}, \mathcal{A}, P)$ is called probability space and is all we need to compute the probability of any event.

However,

- \mathcal{S} is an abstract set, i.e. it contains objects of any kind (faces of a die, faces of a coin, etc.)
- in statistics we deal with data, i.e. numbers s.t. 1.2 kW/h, 10 defective items

Thus, the question: How do we conjugate sample spaces and events to data?

Answer: by the concept of a random variable (r.v.).

Random variables (cont'd)

Random variable: mapping $X : \mathcal{S} \rightarrow \mathbb{R}$ that assigns a real number $X(s)$ to s , for all $s \in \mathcal{S}$.

Example 5 (Single die problem)

When we say “the probability of an odd number equals $1/2$ ”, we are using the r.v.

$$X = \begin{cases} 1 & \text{if } s = \square \bullet \\ 2 & \text{if } s = \square \bullet \bullet \\ \vdots & \\ 6 & \text{if } s = \begin{smallmatrix} \vdots \\ \vdots \\ \vdots \end{smallmatrix} \end{cases}$$

Probability of a random variable

In a triple $(\mathcal{S}, \mathcal{A}, P)$, to each $s \in \mathcal{S}$ there is associated a probability, through P .

Thus, there is a probability associated to each $X(s)$ and, if we have a subset B of reals, using P we can compute the **probability that X takes values in B** .

Formally, let $B \subseteq \mathbb{R}$ then

$$P(X \in B) = P(\{s : X(s) \in B\}).$$

X is continuous if $P(X = x) = 0$ for all $x \in \mathbb{R}$.

X is discrete if $P(X = x) > 0$, for all $x \in \mathbb{R}$ in the range of X .⁶

⁶If X is continuous, its range is uncountable; if X is discrete, its range is countable; X can also be mixed, discrete and continuous.

The probability (density) function (discrete)

Let $\mathcal{X} = \{x_1, x_2, \dots\}$, be the range of X

Probability (density) function (pdf) of X : $p(x) = P(X = x), \forall x \in \mathcal{X}$.

Example 6 (Single die problem)

Let X , s.t. $X = -1$ if the die shows less than three dots, $X = 0$ if it shows three dots and $X = 1$ if it shows more than three dots. We have

$$\begin{aligned} P(X = -1) &= P(\{s : X(s) = -1\}) = P(\{\square, \square\}) \\ &= P(\{\square\}) + P(\{\square\}) = 1/3, \end{aligned}$$

union of disjoint events

and the pdf of X is $p(x) = \begin{cases} 1/3 & \text{if } x = -1 \\ 1/6 & \text{if } x = 0 \\ 1/2 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$



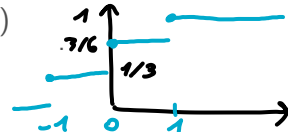
The distribution function

(Cumulative) distribution function (df) of X : $F(x) = P(X \leq x)$,
 $\forall x \in \mathbb{R}$.⁷

Example 7 (Example 6 cont'd)

Considering that

$$F(-1) = P(X \leq -1) = 1/3, \quad F(0) = P(X \leq 0) = 3/6, \quad \text{then}$$

$$F(x) = \begin{cases} 0 & \text{if } x \in (-\infty, -1) \\ 1/3 & \text{if } x \in [-1, 0) \\ 3/6 & \text{if } x \in [0, 1) \\ 1 & \text{if } x \in [1, \infty) \end{cases}$$


The graph shows a step function F(x) on a coordinate system. The x-axis has tick marks at -1, 0, and 1. The y-axis has tick marks at 1/3, 3/6, and 1. The function is 0 for x < -1, jumps to 1/3 at x = -1, jumps to 3/6 at x = 0, and jumps to 1 at x = 1. The function is constant between these points. Handwritten blue annotations highlight the jump points and the corresponding y-values.

Thus F is right-continuous and has jumps at $-1, 0, 1$ and is continuous everywhere.

⁷Yes, F is defined for all $x \in \mathbb{R}$.

Continuous r.v.

For a continuous r.v. if there exists a function $f : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$, s.t.

$$\int_{\mathbb{R}} f(x) dx = 1,$$

and for every $a \leq b$,

$$P(a < X < b) = \int_a^b f(x) dx.$$

then f is the pdf of X .

The df of X : $F(x) = \int_{-\infty}^x f(t) dt, \quad \forall x \in \mathbb{R}.$

pdf and df are related:

$$f(x) = \partial F(x) / \partial x, \text{ at all continuity points of } F.$$

Continuous r.v.

Example 8 (Exponential r.v.)

Let X be the emission time of the first electron from a cathod of a vacuum tube. Under certain physical assumptions it turns out that

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{x \geq 0},$$

provides a ‘good’ approximation to the pdf of X ; $\mathbf{1}_{x \geq 0}$ is an indicator function assuming 1 if $x \geq 0$ and zero otherwise. The df is then

$$F(x) = \int_{-\infty}^x f(t) dt = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Properties of a df

Given a function F how can we be sure it is a df?

The df has the following properties:

- (i) F is nondecreasing;
- (ii) F is continuous from the right;
- (iii) $\lim_{x \rightarrow -\infty} F(x) = 0$;
- (iv) $\lim_{x \rightarrow \infty} F(x) = 1$.

Some further properties:

- (a) $P(X = x) = F(x) - F(x^-)$, where $F(x^-) = \lim_{y \rightarrow x^-} F(y)$
- (b) $P(x < X \leq y) = F(y) - F(x)$
- (c) $P(X > x) = 1 - F(x)$;
- (d) $F(b) - F(a) = P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b)$, for a continuous X .

Example 9 (Exmple 7 cont'd)

$$P(X = 1/2) = F(1/2) - F(1/2^-) = 1/2 - 1/2 = 0$$

$$P(-1 < X \leq 1) = F(1) - F(-1) = 1 - 1/3 = 2/3$$

but also

$$P(-1 < X \leq 1) = P(X \in \{0, 1\}) = 1/6 + 1/2 = 2/3.$$

The quantile function

We need to be careful to define this for discrete rvs 

For a rv X with df F , the quantile function or inverse df is defined by

$$Q(p) = \inf\{x : F(x) \geq p\}, \quad \forall p \in (0, 1).$$


excluded

$Q(1/4)$ is called the first quartile,

$Q(1/2)$ is the second quartile or the median,

$Q(3/4)$ is third quartile. In general

$$\xi_p = Q(p), \quad \text{for any } p \in (0, 1),$$

is called the p th quantile of X .⁸

in practice I take the df draw an horizontal line, and take the leftmost point in the intersection of the line with the df

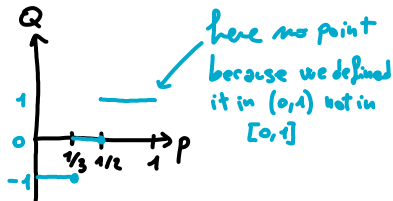
⁸We may assign $Q(0) = -\infty$ and $Q(1) = \infty$.

The quantile function

Example 10 (Example 7 cont'd)

Applying the above definition gives

$$Q(p) = \begin{cases} -1 & \text{if } p \in (0, 1/3] \\ 0 & \text{if } p \in (1/3, 3/6] \\ 1 & \text{if } p \in (3/6, 1) \end{cases}$$



probabilmente sbagliato ci andrebbe (non [

Notable rv's

discrete

- Bernoulli: $f(x) = \theta^x(1 - \theta)^{1-x}$, $x = 0, 1$, $\theta \in [0, 1]$, notation: $\text{Ber}(\theta)$;
- Binomial: $f(x) = \binom{n}{x}\theta^x(1 - \theta)^{n-x}$, $x = 0, 1, \dots, n$, $\theta \in [0, 1]$, notation: $\text{Bin}(n, p)$;
- Negative Binomial: $f(x) = \binom{x+r-1}{x}\theta^r(1 - \theta)^x$, $x, r \in \mathbb{Z}_{\geq 0}$, notation: $\text{NegBin}(r, p)$; If $r = 1$ it's called geometrical dist.
- Poisson: $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $y \in \mathbb{Z}_{\geq 0}$, $\lambda > 0$, notation $\text{Poi}(\lambda)$;

Continuous

- Gaussian: $f(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}} / (\sqrt{2\pi}\sigma)$, $x, \mu \in \mathbb{R}$, $\sigma^2 > 0$, notation: $N(\mu, \sigma^2)$;
- Exponential: $f(x) = \lambda e^{-\lambda x}$, $y \in \mathbb{R}_{\geq 0}$, $\lambda > 0$, notation: $\text{Exp}(\lambda)$;
- Gamma: $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, $y \in \mathbb{R}_{\geq 0}$, $\alpha > 0, \lambda > 0$, notation: $\text{Ga}(\alpha, \lambda)$;
- Weibull: $f(x) = \frac{\alpha}{\beta} x^{\alpha-1} e^{-\frac{x^\alpha}{\beta}}$, $y \in \mathbb{R}_{\geq 0}$, $\alpha > 0, \beta > 0$, notation: $\text{Wei}(\alpha, \beta)$;
- Uniform: $f(x) = (b - a)^{-1} \mathbf{1}_{[a,b]}$, $a, b \in \mathbb{R}$, $a < b$ notation: $\text{Unif}(a, b)$;

We write $X \sim F$ to say that 'X is distributed as F'.

- Chi-squared distribution: If $Z_i \sim N(0, 1)$, $i = 1, \dots, n$, and Z_i 's are independent⁹ then

$$Z_1^2 + \dots + Z_n^2 \sim \chi_n^2,$$

n is called degrees of freedom;

- t distribution: if $Z \sim N(0, 1)$ and $U \sim \chi_\nu^2$, with Z, U independent, then

$$Z/\sqrt{U/\nu} \sim t_\nu,$$

, ν is called degrees of freedom.

- F distribution: if $U_1 \sim \chi_{n_1}^2$ and $U_2 \sim \chi_{n_2}^2$, with U_1, U_2 independent, then

$$(U_1/n_1)/(U_2/n_2) \sim F_{n_1, n_2},$$

and n_1, n_2 are the numerator and denominator degrees of freedom, resp.

⁹More on independence of rv's later

Moments

For $g : \mathbb{R} \rightarrow \mathbb{R}$, the expectation of $g(X)$ is

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx,$$

provided $\int_{-\infty}^{\infty} |g(x)|f(x)dx$ exists and is finite.

Examples

- $g(x) = x$: $E(X) = \mu_X$ the expectation of X ;
- $g(x) = (x - c)^n$, $c \in \mathbb{R}$, $n \in \mathbb{N}$: n th moment of X about c

$$E[(X - c)^n] = \int_{-\infty}^{\infty} (x - c)^n f(x)dx,$$

provided the integral exists;

- $g(x) = (x - E(X))^n$: n th central moment; for $n = 2$, we get the variance of X , denoted $\text{var}(X) = \sigma_X^2$.

Transformation of a r.v.

Applying a function g to a r.v. X , leads to another r.v. $Y = g(X)$.

If X is discrete, the Y is discrete and

$$f_Y(y) = P(Y = y) = P(\{s : g(X(s)) = y\}).$$

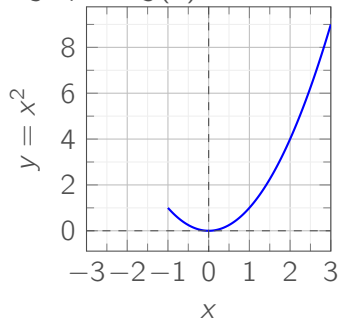
If X is continuous the procedure is more difficult, we have to:

- find $B_y = \{x : g(x) \leq y\}$, for each y in the range of Y ;
- find the df
$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(B_y) = \int_{B_y} f(x)dx;$$
- take the derivative, i.e. $f_Y(y) = F'_Y(y)$.

Transformation of a rv

Example 11

Let $X \sim \text{Unif}(-1, 3)$, we compute the pdf of $Y = X^2$. Consider the graph of $g(x) = x^2$.



Now $Y \in [0, 9]$, for $x \in [-1, 1]$, $g(x) \in [0, 1]$, whereas for $x > 1$ $g(x)$ is bijective. We have two cases:

- (1) $0 \leq y \leq 1$: $B_y = (-\sqrt{y}, \sqrt{y})$ and
$$F_Y(y) = (1/4) \int_{B_y} \mathbf{1}_{[-1,3]} dx = \sqrt{y}/2;$$
- (2) $y > 1$: $B_y = (-1, \sqrt{y})$,
$$F_Y(y) = (\sqrt{y} + 1)/4.$$

The pdf can be found by differentiating F_Y , taking care of the two cases.

Example 12 (Example 11 cont'd)

Let's compute $E(Y)$ and $\text{var} Y$. We have that

$$f_Y(y) = \begin{cases} 1/(4\sqrt{y}) & \text{if } 0 < y \leq 1 \\ 1/(8\sqrt{y}) & \text{if } 1 < y \leq 9. \end{cases}$$

Then

$$\begin{aligned} E(Y) &= \int_0^9 y f_Y(y) dy = \int_0^1 y/(4\sqrt{y}) dy + \int_1^9 y/(8\sqrt{y}) dy \\ &= \int_0^1 \sqrt{y}/4 dy + \int_1^9 \sqrt{y}/8 dy = 7/3. \end{aligned}$$

$$\begin{aligned} E(Y^2) &= \int_0^9 y^2 f_Y(y) dy = \int_0^1 y^2/(4\sqrt{y}) dy + \int_1^9 y^2/(8\sqrt{y}) dy \\ &= \int_0^1 y^{3/2}/4 dy + \int_1^9 y^{3/2}/8 dy = 61/5. \end{aligned}$$

Thus $\text{var}(Y) = E(Y^2) - E(Y)^2 = 61/5 - 49/9 = 794/45$.

Inverse transform

Suppose Y is a continuous rv with distribution F_Y .

It can be shown that F_Y is continuous and bijective with inverse $F^{-1} = Q$.

Furthermore, if $X = \text{Unif}(0, 1)$, then $Q(X) \sim F_Y$.

This fact is useful when we want to draw random values from F_Y .
Indeed, if we

(a) draw a number p uniformly in $(0, 1)$

(b) set $y = Q(p)$,

y is a random value from F_Y . This is known as the **inverse transform sampling** method.