Mathematics

Handbook of Statistics

Author:
David Silva Sanmartín

July 24, 2020

Contents

Part I

Theory

Chapter 1

Probability spaces

1.1 Definition and basic properties

Definition 1. Random Experiment: an experiment whose outcomes are determined only by chance factors.

Definition 2. The description of a random experiment starts by identifying the set of all possible outcomes of the experiment, which we call **sample space**, and we designate by Ω.

- When Ω is finite or countable (*numerable*), we say that the probability model is **discrete**.
- When Ω is uncountable, we say that the probability model is **continuous**.

Definition 3. Let Ω be a set and $\mathcal{F}$ a σ-algebra on Ω. We say that $(\Omega, \mathcal{F})$ is a **measurable space**. Any subset belonging to $\mathcal{F}$ is called an **event**.

- When Ω is finite or countable, we will usually consider $\mathcal{F} = \mathcal{P}(\Omega)$.
- When Ω is uncountable, with $\Omega \subseteq \mathbb{R}^k$, we will usually consider $\mathcal{F} = \mathbb{B}_\Omega^k$.

Definition 4. Given a measurable space $(\Omega, \mathcal{F})$, a **probability**, or **probability measure**, is a function $\mathrm{P} : \mathcal{F} \to \mathbb{R}$ such that:

i) $\mathrm{P}(A) \geq 0$ for every $A \in \mathcal{F}$.

ii) For every countable collection of events $\{A_n\} \subset \mathcal{F}$, which are pairwise disjoint, we have that:

$$\mathrm{P}\left(\cup_n A_n\right) = \sum_n \mathrm{P}(A_n)$$

iii) $\mathrm{P}(\Omega) = 1$

If P is a probability in $(\Omega, \mathcal{F})$, we call the triad $(\Omega, \mathcal{F}, \mathrm{P})$ a **probability space**. For every event $A \in \mathcal{F}$, we call $\mathrm{P}(A)$ the **probability of** A.

Proposition 5. *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, and $A, B \in \mathcal{F}$. Then:*

$$\mathrm{P}(A - B) = \mathrm{P}(A) - \mathrm{P}(A \cap B)$$

If $B \subseteq A$, then

$$\mathrm{P}(A - B) = \mathrm{P}(A) - \mathrm{P}(B)$$

and, thus, $\mathrm{P}(B) \leq \mathrm{P}(A)$. In particular, it is $\mathrm{P}(A) \leq 1$ for every $A \in \mathcal{F}$, and:

$$\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$$

And from this last equality we deduce that $\mathrm{P}(\emptyset) = 0$.

Proposition 6. *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, and $A_1, \ldots, A_n \in \mathcal{F}$. Then,*

$$\mathrm{P}(A_1 \cup A_2) = \mathrm{P}(A_1) + \mathrm{P}(A_2) - \mathrm{P}(A_1 \cap A_2)$$

*and, more generally, the **inclusion-exclusion formula** holds:*

$$\mathrm{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathrm{P}(A_i) - \sum_{1 \leq i \leq j \leq n} \mathrm{P}(A_i \cap A_j) + \sum_{1 \leq i \leq j \leq k \leq n} \mathrm{P}(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} \mathrm{P}(A_1 \cap A_2 \cap \cdots A_n)$$

1.2 Conditional probability

In a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, the fact of knowing that the outcome of the random experiment verifies a particular event B, affects the probability of all events.

Definition 7. Let $B \in \mathcal{F}$ be an event such that $\mathrm{P}(B) > 0$. For each $A \in \mathcal{F}$, we define the **conditional probability of** A **given** B as:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

The knowledge that the event B has been verified transforms the original probability space into a new one, $(\Omega, \mathcal{F}, \mathrm{P}(\cdot \mid B))$. Because we also know that the event B^c has not happened, we can consider a more relevant space, $(B, \mathcal{F}_B, \mathrm{P}_B)$, in which:

$$\mathcal{F}_B = \{A \cap B \mid A \in \mathcal{F}\}$$

is the new σ-algebra of events. We assign each of the new events, $C \in \mathcal{F}_B$, the probability:

$$\mathrm{P}_B(C) = \frac{\mathrm{P}(C)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}, \text{ if } C = A \cap B \text{ for some } A \in \mathcal{F}$$

Proposition 8. *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. If $B \in \mathcal{F}$ is an event such that $\mathrm{P}(B) > 0$, then both $(\Omega, \mathcal{F}, \mathrm{P}(\cdot \mid B))$ and $(B, \mathcal{F}_B, \mathrm{P}_B)$ are probability spaces.*

Proposition 9 (Product rule or Chain rule)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, and $A_1, \ldots, A_n \in \mathcal{F}$ events such that $\mathrm{P}(\cap_{i=1}^{n-1} A_i) > 0$. Then,*

$$\mathrm{P}(\cap_{i=1}^{n} A_i) = \mathrm{P}(A_1)\mathrm{P}(A_2 \mid A_1)\mathrm{P}(A_3 \mid A_2 \cap A_1) \cdots \mathrm{P}(A_n \mid \cap_{i=1}^{n-1} A_i)$$

Proof. By iteratively applying the definition of conditional probability, $\mathrm{P}(A \cap B) = \mathrm{P}(B)\mathrm{P}(A \mid B)$, we have that:

$$\begin{aligned} \mathrm{P}(\cap_{i=1}^{n} A_i) &= \mathrm{P}(\cap_{i=1}^{n-1} A_i)\mathrm{P}(A_n \mid \cap_{i=1}^{n-1}) = \\ &= \mathrm{P}(\cap_{i=1}^{n-2} A_i)\mathrm{P}(A_{n-1} \mid \cap_{i=1}^{n-2} A_i)\mathrm{P}(A_n \mid \cap_{i=1}^{n-1}) = \\ &= \cdots \end{aligned}$$

□

Proposition 10 (Law or formula of total probability)**.** *If $B_n \subset \mathcal{F}$ is a finite or countably infinite partition of the sample space (in other words, a set of pairwise disjoint events whose union is Ω) and $\mathrm{P}(B_n) > 0$ for each n, then for any event $A \in \mathcal{F}$:*

$$\mathrm{P}(A) = \sum_n \mathrm{P}(B_n)\mathrm{P}(A \cap B_n)$$

Proof.

$$\mathrm{P}(A) = \mathrm{P}(A \cap \Omega) = \mathrm{P}(A \cap (\cup_n B_n)) = \mathrm{P}(\cup_n (A \cap B_n)) = \sum_n \mathrm{P}(A \cap B_n) =$$

$$= \sum_n \mathrm{P}(B_n)\mathrm{P}(A \cap B_n)$$

□

Proposition 11 (Bayes' formula)**.** *If A and B are events in the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ such that $\mathrm{P}(B) > 0$, then:*

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A)\,\mathrm{P}(B \mid A)}{\mathrm{P}(B)}$$

Proof.

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A)\,\mathrm{P}(B \mid A)}{\mathrm{P}(B)}$$

□

1.3 Independence

Definition 12. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. $\{A_i\}_{i\in I}$ are **independent events** if:

$$\mathrm{P}(\cap_{i\in F} A_i) = \prod_{i\in F} \mathrm{P}(A_i)$$

for every finite subset $F \subset I$.

Definition 13. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $\{\mathcal{C}_i\}_{i\in I}$ families of events, where $\mathcal{C}_i \subset \mathcal{F}$ for each $i \in I$. $\{\mathcal{C}_i\}_{i\in I}$ are said to be **independent** if the events $\{A_i\}_{i\in I}$ are independent for every election of $A_i \in \mathcal{C}_i$ for every $i \in I$.

Example 14. Let $\mathbb{B}^k$ be the Borel σ-algebra in $\mathbb{R}^k$. For every $B \in \mathbb{B}^k$ we define:

$$\mathrm{P}(B) = \frac{\lambda_k(B \cap A)}{\lambda_k(A)}$$

where $A \in \mathbb{B}^k$ with finite volume $\lambda_k(A) > 0$. Then,

i) P is a probability in $(\mathbb{R}^k, \mathbb{B}^k)$, which is called the **uniform distribution** on A.

ii) If $A = A_1 \times A_2$ with $A_1 \in \mathbb{B}^r$, $A_2 \in \mathbb{B}^s$ and $r + s = k$, then

$$\{B_1 \times A_2 \mid B_1 \in \mathbb{B}^r\} \text{ and } \{A_1 \times B_2 \mid B_2 \in \mathbb{B}^s\}$$

are families of independent events.

iii) If $A = A_1 \times \cdots \times A_k$, with $A_i, \ldots, A_k \in \mathbb{B}$, then the families of events $\mathcal{C}_i = \{A_1 \times \cdots \times B_i \times \cdots \times A_k \mid B_i \in \mathbb{B}\}$ are independent.

Chapter 2

Probabilities in $\mathbb{R}$

2.1 Distribution functions

In this chapter, we will consider probability spaces of the form $(\mathbb{R}, \mathbb{B}, \mathrm{P})$, where the only variable is the probability function P. It is posible to characterize any probability function P in $\mathbb{R}$ via a real-valued real function:

Definition 15. A **distribution function** is a function $F : \mathbb{R} \to [0, 1]$ such that:

1. F increases in $\mathbb{R}$: $F(x_1) \leq F(x_2)$ for all $x_1 < x_2$.
2. F is right-continuous: for every $x \in \mathbb{R}$, $F(x) = \lim_{y \to x, y > x} F(y)$.
3. $\lim_{y \to -\infty} F(x) = 0$ and $\lim_{y \to \infty} F(x) = 1$.

Condition 1 guarantees the existence of both one-sided limits:

$$F(x^+) = \lim_{y \to x, y > x} F(y) \text{ and } F(x^-) = \lim_{y \to x, y < x} F(y)$$

From condition 2, we have that $F(x^+) = F(x)$. However, at some points, we could have that $F(x^-) \neq F(x)$. In this case, F presents a jump discontinuity at x.

TODO Examples...

Any probability fnuction P in $(\mathbb{R}, \mathbb{B})$ determines a distribution function:

Proposition 16. *Let* P *be a probability function in* $(\mathbb{R}, \mathbb{B})$*. Then, the function* F *defined as follows:*

$$F(x) = \mathrm{P}((-\infty, x])$$

is a distribution function.

Chapter 3

Random variables

3.1 Random variables and expectations

Definition 17. Random Experiment: an experiment whose outcomes are determined only by chance factors.

Definition 18. Sample Space: the set of all possible outcomes of a random experiment.

Definition 19. Event: the collection of none, one, or more than one outcomes from a sample space.

Definition 20. Random Variable: a variable whose numerical values are determined by chance factors. It is a function from the sample space to a set of real numbers.

Given a sample space $\Omega = \{\omega_1, \ldots, \omega_n\}$ with probability function P and a random variable X with range $\mathbb{X} = \{x_1, \ldots x_m\}$, we can define a probability function P_X on $\mathbb{X}$ in the following way: We will observe $X = x_i$ if and only if the outcome of the random experiment is an $\omega_i \in \Omega$ such that $X(\omega_j) = x_i$. Thus,

$$P_X(X = x_i) = P(\{\omega_j \in \Omega \mid X(\omega_j) = x_i\})$$

The function P_X is an *induced* probability function on $\mathbb{X}$, defined in terms of the orginal function P. We will write $P(X = x_i)$ rather than $P_X(X = x_i)$.

Definition 21. Discrete Random Variable: if the set of all possible values of a random variable X is countable, then X is called a *discrete random variable*.

Definition 22. Probability Mass Function (pmf): let R be the set of all possible values of a discrete random variable X, and $f(k) = P(X = k)$ for each k in R. Then $f(k)$ is called the *probability mass function* of X.

Definition 23. Continuous Random Variable: if the set of all possible values of X an interval or union of two or more nonoverlapping intervals in $\mathbb{R}$, then X is called a *continuous random variable*.

Definition 24. Probability Density Function (pdf): any real valued function $f(x)$ that satisfies the following requirements is called a *probability density function*:

$$f(x) \geq 0 \text{ for all } x, \text{ and } \int_{-\infty}^{\infty} f(x)\,dx = 1$$

Definition 25. Cumulative Distribution function (cdf): the *cumulative distribution function* of a random variable X is defined by

$$F(x) = P(X \leq x)$$

For a continuous random variable X with the probability density function $f(x)$,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt, \ \forall x$$

For a discrete random variable X, the *cdf* is defined by

$$F(k) = P(X \leq k) = \sum_{i=-\infty}^{k} P(X = i)$$

The *pdf* (or *pmf*) contains the same information as the *cdf*.

If the distribution of a random variable X depends on a parameter θ, the the *pdf* or *pmf* of X is usually expressed as $f(x \mid \theta)$, and the *cdf* is written as $F(x \mid \theta)$

Theorem 26. *A function $F(x)$ is a cdf if and only if the following three conditions hold:*

a) $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$.
b) $F(x)$ *is a nondecreasing function of x.*
c) $F(x)$ *is right-continuous; that is, for every number x_0,* $\lim_{x\to x_0^+} F(x) = F(x_0)$

Definition 27. Expectation: if X is a continuous random variable with the *pdf* $f(x)$, then the expectation of $g(X)$, where g is a real valued function, is defined by

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

If X is a discrete random variable, then

$$E(g(X)) = \sum_k g(x)P(X = k)$$

where the sum is over all possible values of X. Thus, $E(g(X))$ is the weighted average of the possible values of $g(X)$, each weighted by its probability.

Theorem 28. *Let X be a random variable and let a, b and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,*

a) *$E(ag_1(X) + bg_2(X) + c) = aEg_1(X) + bEg_2(X) + c$.*
b) *If $g_1(x) \geq 0\ \forall x$, then $Eg_1(X) \geq 0$.*
c) *If $g_1(x) \geq g_2(x)\ \forall x$, then $Eg_1(X) \geq Eg_2(X)$.*
d) *If $a \leq g_1(x) \leq b\ \forall x$, then $a \leq Eg_1(X) \leq b$.*

3.2 Measures of Central Tendency

Definition 29. Mean: the *mean* of a random variable X is usually denoted by μ. For a discrete random variable X, it is defined by

$$\mu = E(X) = \sum_k P(X = k)$$

where the sum is over all possible values of X. For a continuous random variable X with probability density function $f(x)$, the *mean* is defined by

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

Definition 30. Median: the *median* of a random variable X is the value such that 50% of the possible values of X are less than or equal to that value. For a discrete distribution, median is not well defined, and it need not be unique.

Definition 31. Mode: the most probable value of the random variable.

3.3 Moments

Definition 32. Moments about the origin (Raw Moments): for each integer n, the nth *moment* of X (or $F(X)$), μ'_n, is

$$\mu'_n = EX^n$$

For a continuous random variable this is:

$$\mu'_n = EX^n = \int_{-\infty}^{\infty} x^n f(x)\, dx$$

We can see that the mean is $\mu = \mu'_1 = EX$.

Definition 33. Moments about the mean (Central Moments): for each integer n, the nth *central moment* of X (or $F(X)$), μ_n, is

$$\mu_n = E(X - \mu)^n$$

Definition 34. Variance: the *variance* of a random variable X is its second central moment:

$$\sigma^2 = VarX = E(X - EX)^2$$

An alternative formula for the variance is given by

$$VarX = EX^2 - (EX)^2$$

Definition 35. Standard deviation: the *standard deviation* of a random variable X is the square root of $VarX$:

$$\sigma = \sqrt{VarX}$$

The variance gives a measure of the degree of spread of a distribution around its mean. Larger values mean X is more variable. At the extreme, if $VarX = E(X - EX)^2 = 0$, then X is equal to EX with probability 1, and there is no variation in X.

The standard deviation has the same qualitative interpretation: small values mean X is very likely to be close to EX, and large values mean X is very variable. The standard deviation is easier to interpret in that the measurement unit on the standard deviation is the same as that for the original variable X.

Theorem 36. *If X is a random variable with finite variance, then for any constants a and b:*

$$Var(aX + b) = a^2 VarX$$

Definition 37.

Part II

Discrete distributions

Chapter 4

Discrete uniform distribution

4.1 Description

Used to model experimental outcomes which are "equally likely".

4.1.1 Probability mass function

$$P(X = k) = \frac{1}{N}, \quad k = 1, \ldots, N$$

4.1.2 Cumulative distribution function

$$P(X \leq k) = \frac{k}{N}, \quad k = 1, \ldots, N$$

4.2 Moments

Mean	$\frac{N+1}{2}$
Variance	$\frac{(N-1)(N+1)}{2}$

Chapter 5

Binomial distribution

5.1 Description

A binomial experiment involves n independent and identical trial such that each trial can result into one of the two possible outcomes: success of failure. If p is the probability of observing success in each trial, then the number of successes X that can be observed out of these n trials is referred to as the **binomial random variable with n trials and success probability** p, or $B(n, p)$.

Binomial distribution is often used to estimate or determine the proportion of individuals with a particular attribute in a large population. Suppose that a random sample of n units is drawn by sampling with replacement from a finite population or by sampling without replacement from a large population. The number of units that contain the attribute of interest in the sample follows a normal distribution.

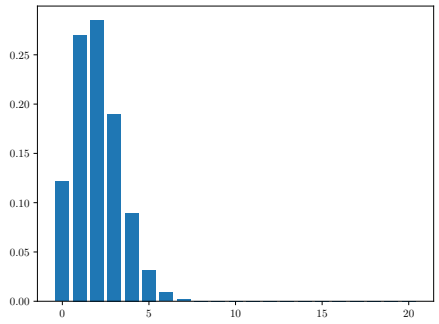
If the sample was drawn without replacement from a small finite population, the hypergeometric distribution should be used instead of the binomial.

5.1.1 Probability mass function

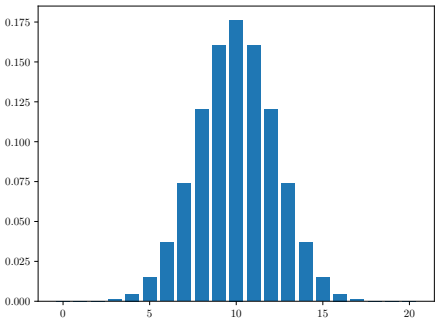
The probability of observing k successes out of n trials is given by the following probability mass function

$$P(X = k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

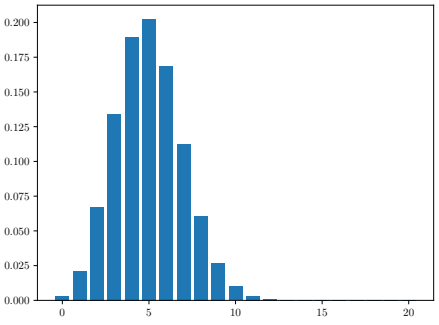
Binomial's pmf is right-skewed when $p < 0.5$, left-skewed when $p > 0.5$ and symmetric when $p = 0.5$.



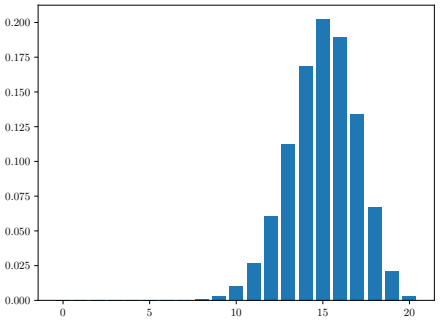
(a) $B(20, 0.1)$



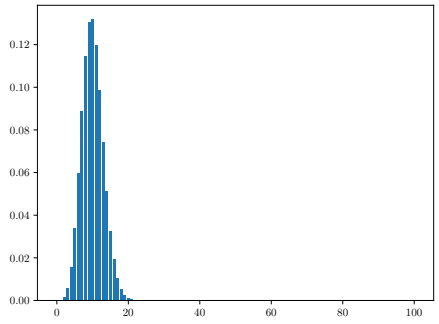
(b) $B(20, 0.5)$



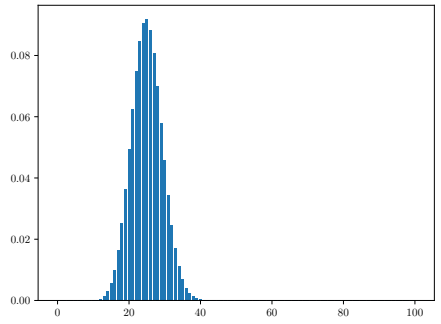
(c) $B(20, 0.25)$



(d) $B(20, 0.75)$



(e) $B(100, 0.1)$



(f) $B(100, 0.25)$

Figure 5.1: Binomial distribution

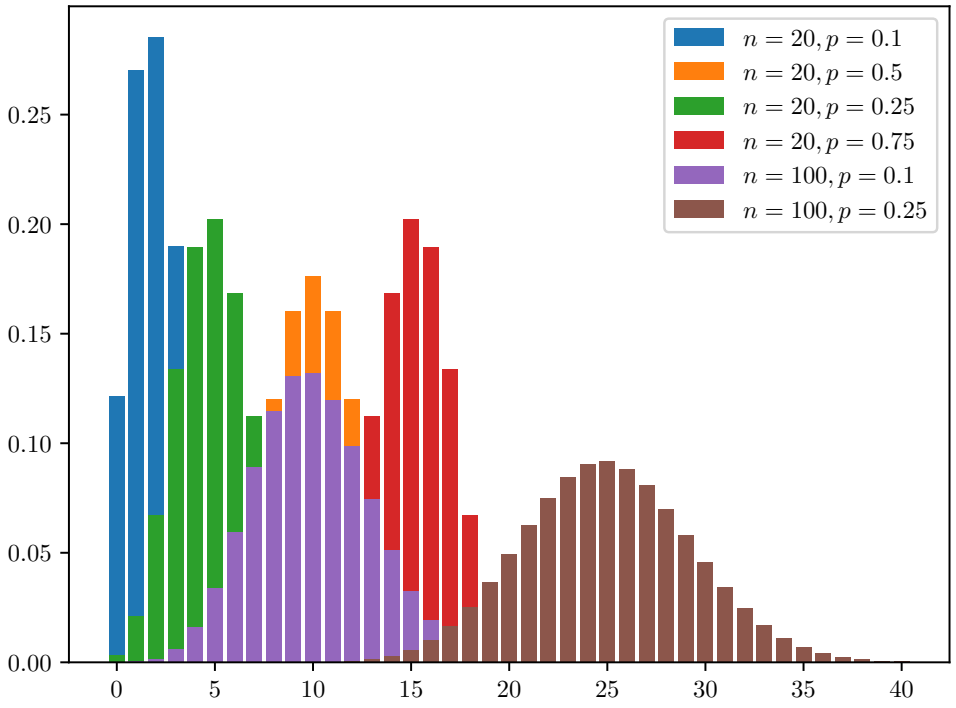
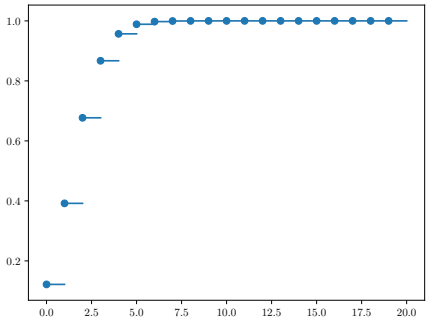


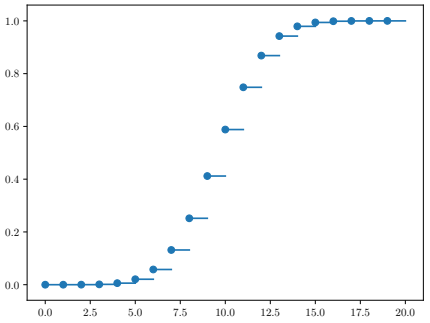
Figure 5.2: $P(X = k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$

5.1.2 Cumulative distribution function

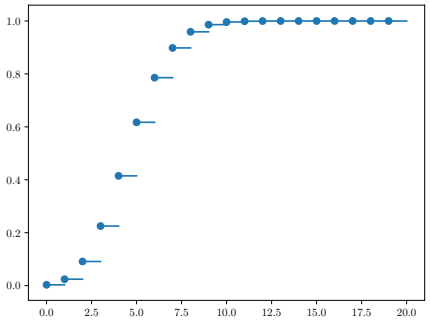
$$P(X \leq k \mid n, p) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}, \quad k = 0, 1, \ldots, n$$



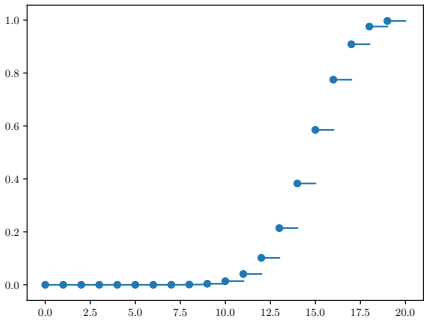
(a) $B(20, 0.1)$



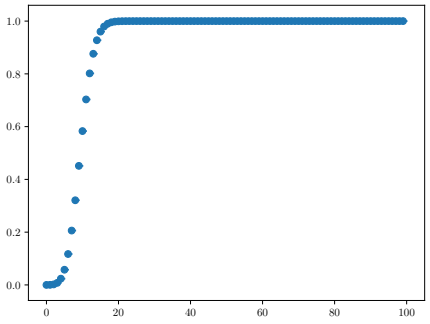
(b) $B(20, 0.5)$



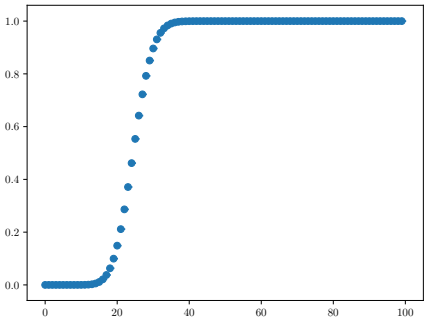
(c) $B(20, 0.25)$



(d) $B(20, 0.75)$



(e) $B(100, 0.1)$



(f) $B(100, 0.25)$

Figure 5.3: Binomial distribution

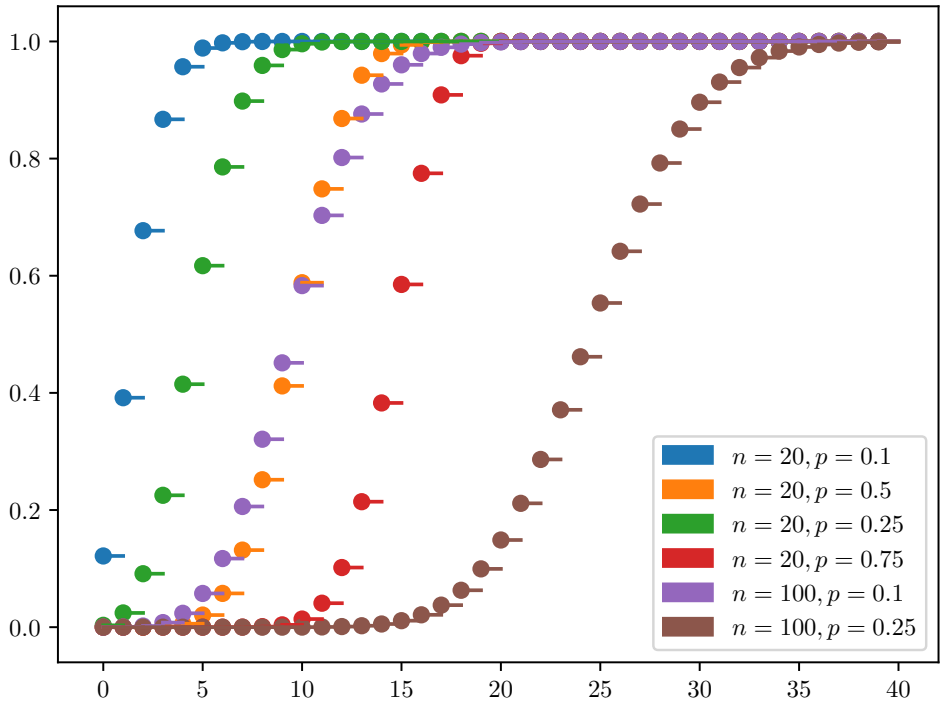


Figure 5.4: $P(X \leq k \mid n, p) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}, \quad k = 0, 1, \ldots, n$

5.2 Moments

Mean	np
Variance	$np(1-p)$

5.3 Properties

1. Let $X_1, \ldots, X_m$ be independent random variables with $X_i \sim B(n_i, p)$, $i = 1, 2, \ldots, m$. Then,

$$\sum_{i=1}^{m} X_i \sim B\left(\sum_{i=1}^{m} n_i, p\right)$$

2. Let $X_1, \ldots, X_m$ be independent Bernoulli(p) random variables with success probability p. That is: $P(X_i = 1) = p,\ P(X_i = 0) = 1 - p,\ i = 1, \ldots, n$. Then,
$$\sum_{i=1}^{m} X_i \sim B\left(n, p\right)$$

5.4 Examples

Example 38. A fair die is rolled n times.

- The probability of obtaining exactly one 6 is $n\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{n-1}$.
- The probability of obtaining no 6 is $\left(\frac{5}{6}\right)^{n}$.
- The probability of obtaining at least one 6 is $1 - \left(\frac{5}{6}\right)^{n}$.
- The number of trials needed for the probability of at least one 6 to be $\geq \frac{1}{2}$ is given by the smallest integer n such that $1 - \left(\frac{5}{6}\right)^{n} \geq \frac{1}{2}$, so that $n \geq \frac{\log 2}{\log 1.2} \approx 3.8$.

Chapter 6

Poisson distribution

6.1 Description

The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

For instance, an individual keeping track of the amount of mail they receive each day may notice that they receive an average number of 4 letters per day. If receiving any particular piece of mail does not affect the arrival times of future pieces of mail, i.e., if pieces of mail from a wide range of sources arrive independently of one another, then a reasonable assumption is that the number of pieces of mail received in a day obeys a Poisson distribution. Other examples that may follow a Poisson distribution include the number of phone calls received by a call center per hour, the number of decay events per second from a radioactive source, the number of visits to a website and the number of typographical errors per page in a book.

The Poisson distribution is an appropriate model if the following assumptions are true:

- k is the number of times an event occurs in an interval and k can take values 0, 1, 2, ...
- The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
- The average rate at which events occur is constant.
- Two events cannot occur at exactly the same instant; instead, at each very

small sub-interval exactly one event either occurs or does not occur.

The Poisson distribution can also be developed as a limiting distribution of the binomial, in which $n \to \infty$ and $p \to 0$ so that np remains a constant. In other words, for large n and small p, the binomial distribution can be approximated by the Poisson distribution with mean $\lambda = np$. If the sample was drawn without replacement from a small finite population, the hypergeometric distribution should be used instead of the binomial.

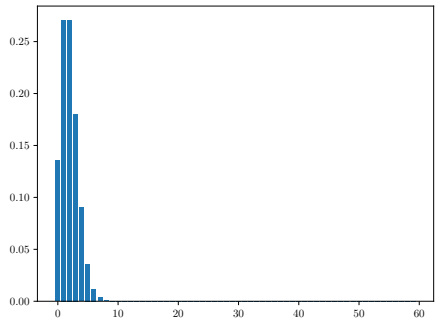
If the number of events per unit time follows a Poisson distribution, then the amount of time between events follows the exponential distribution.

6.1.1 Probability mass function

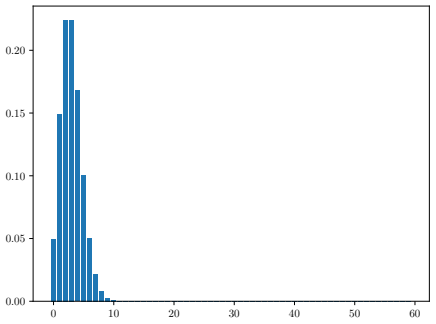
Let X denote the number of events in a unit interval of time or in a unit distance. Then, X is called the Poisson random variable with mean number of events λ in a unit interval of time. The probability mass function of a Poisson distribution with mean λ is given by

$$f(k \mid \lambda) = P(X = k \mid \lambda) = \frac{e^{-\lambda}\lambda^k}{k!},\ k = 0, 1, 2, \ldots$$

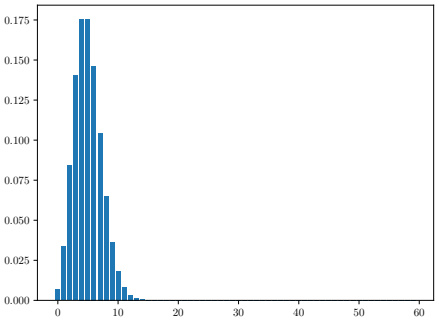
The Poisson probability mass function is right-skewed, because it is inhibited by the zero occurrence barrier (there can not happen less than 0 events) on the left and it is unlimited on the other side. The degree of skewness decreases as λ increases. As λ becomes bigger, the graph looks more like a normal distribution.



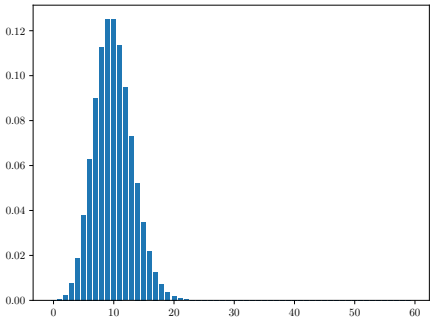
(a) $\lambda = 2$



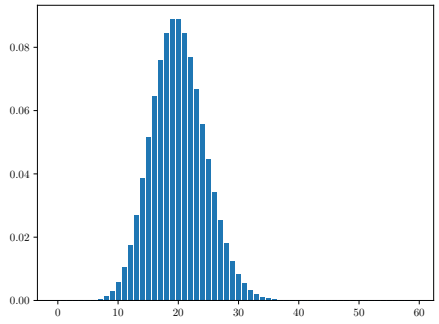
(b) $\lambda = 3$



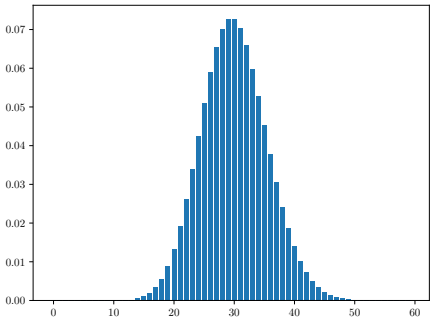
(c) $\lambda = 5$



(d) $\lambda = 10$



(e) $\lambda = 20$



(f) $\lambda = 30$

Figure 6.1: Poisson distribution

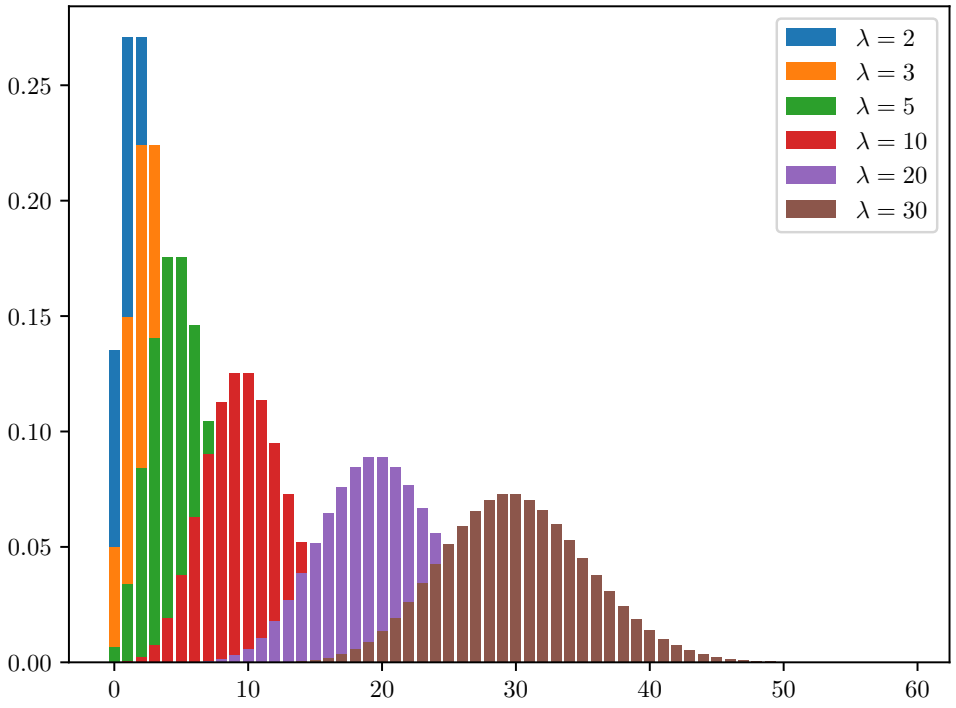
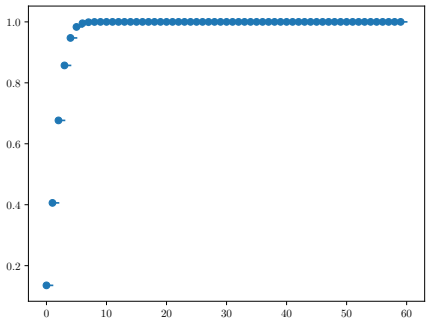


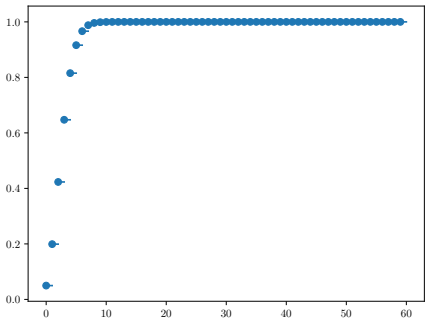
Figure 6.2: $f(k \mid \lambda) = P(X = k \mid \lambda) = \frac{e^{-\lambda}\lambda^k}{k!},\ k = 0, 1, 2, \ldots$

6.1.2 Cumulative distribution function

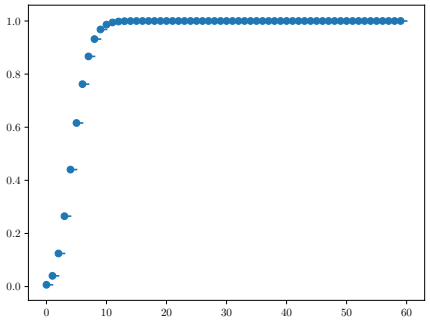
$$F(k \mid \lambda) = P(X = k \leq \lambda) = \sum_{i=0}^{k} \frac{e^{-\lambda}\lambda^i}{i!},\ k = 0, 1, 2, \ldots$$



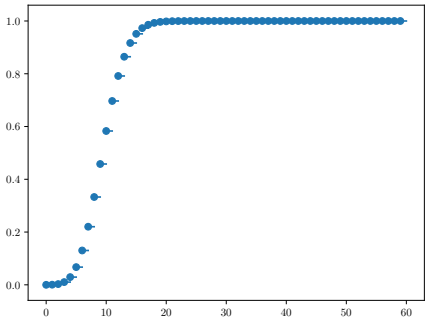
(a) $\lambda = 2$



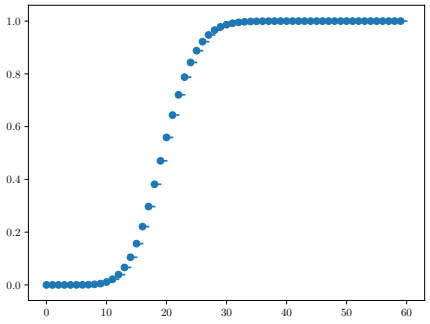
(b) $\lambda = 3$



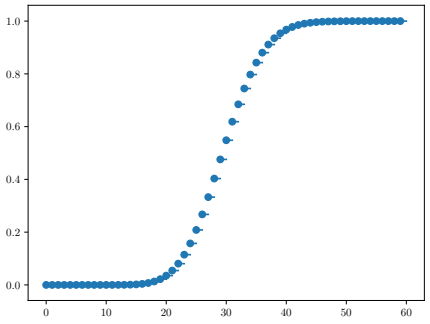
(c) $\lambda = 5$



(d) $\lambda = 10$



(e) $\lambda = 20$



(f) $\lambda = 30$

Figure 6.3: Poisson distribution

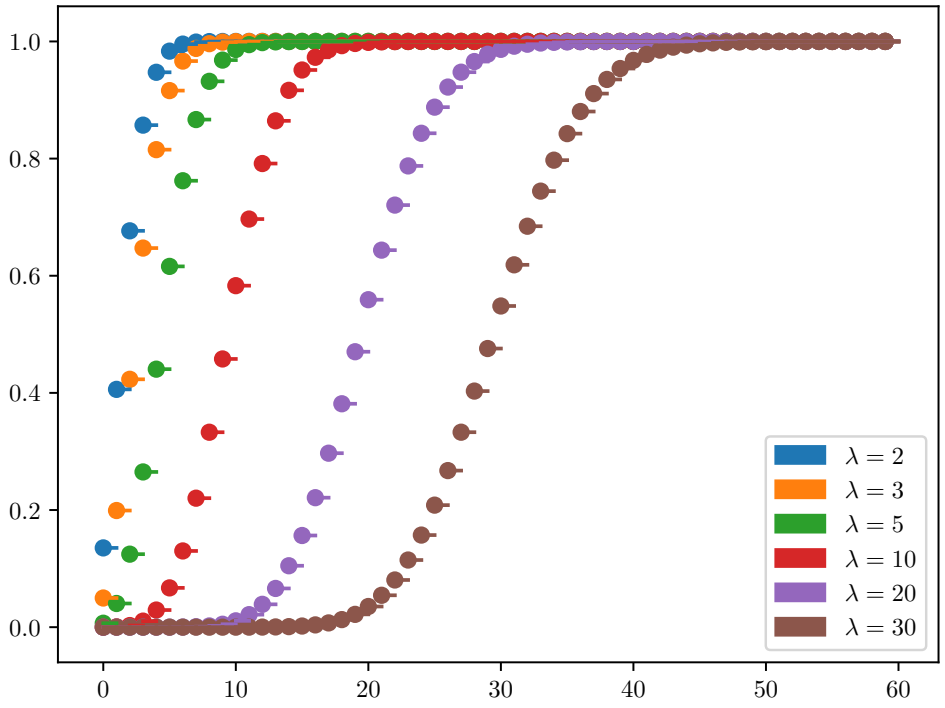


Figure 6.4: $F(k \mid \lambda) = P(X = k \leq \lambda) = \sum_{i=0}^{k} \frac{e^{-\lambda}\lambda^i}{i!},\ k = 0, 1, 2, \ldots$

6.2 Moments

Mean	λ
Variance	λ

6.3 Examples

Example 39 (Approximation of the binomial to a Poisson)**.** (Taken from [3])

Suppose we have a website. Each person that visits the website has some probability of clicking an ad. The binomial distribution can help us calculate the probability of successful events (clicks). A binomial random variable is the number of successes

(x) in n repeated trials. We assume the probability of success p is constant over each trial.

We are interested in knowing the number of people that will click an ad per week. Let's assume we have the stats for a year: a total of 59000 poeple visited the website, and out of them, 888 people clicked an ad. Then, $n = 59000/12 = 1134$. The number of people who clicked an ad per week is $888/52 = 17$. The probability of success is $p = 888/59000 = 0.015$.

At this point we can calculate, for example, the probability that next week exactly 20 people click an add. We use the binomial pmf for that:

$$P(X = 20) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{1134}{20}\ 0.015^2 0\ (1 - 0.015)^{1134-20} = 0.06962$$

Other values are as follows:

x	Binomial $P(X = x)$
10	0.02250
17	0.09701
20	0.06962
30	0.00121
40	< 0.000001

Note that the **expected value** or **mean**, 17 (which is the average number of successes per week we calculated from the raw data), has the highest probability of happening.

Part III

Distributions in $\mathbb{R}$

Chapter 7

Exponential distribution

7.1 Description

The exponential distribution is the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate. It is a particular case of the gamma distribution. It is the continuous analogue of the geometric distribution, and it has the key property of being memoryless. In addition to being used for the analysis of Poisson point processes it is found in various other contexts.

TODO: proof of memorylessness (see MIT notes)

7.1.1 Probability mass function

$$f(x;\lambda) = \begin{cases} 0 & \text{if } x < 0 \\ \lambda e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

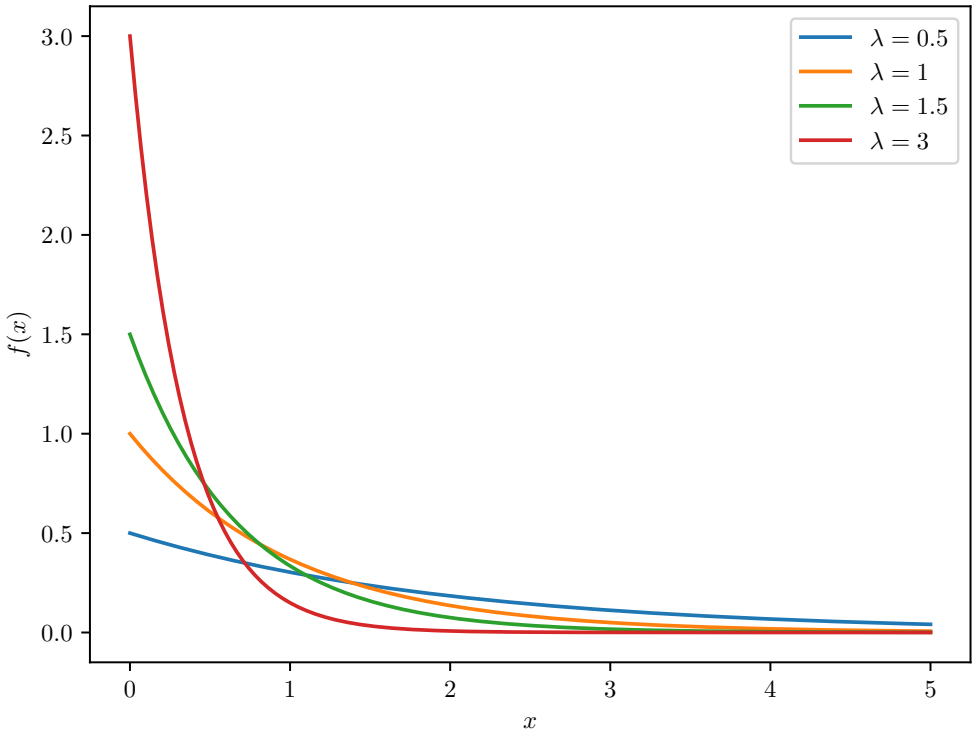


Figure 7.1: $f(x;\lambda) = \lambda e^{-\lambda x}$ if $x \geq 0$

7.1.2 Cumulative distribution function

$$F(x;\lambda) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

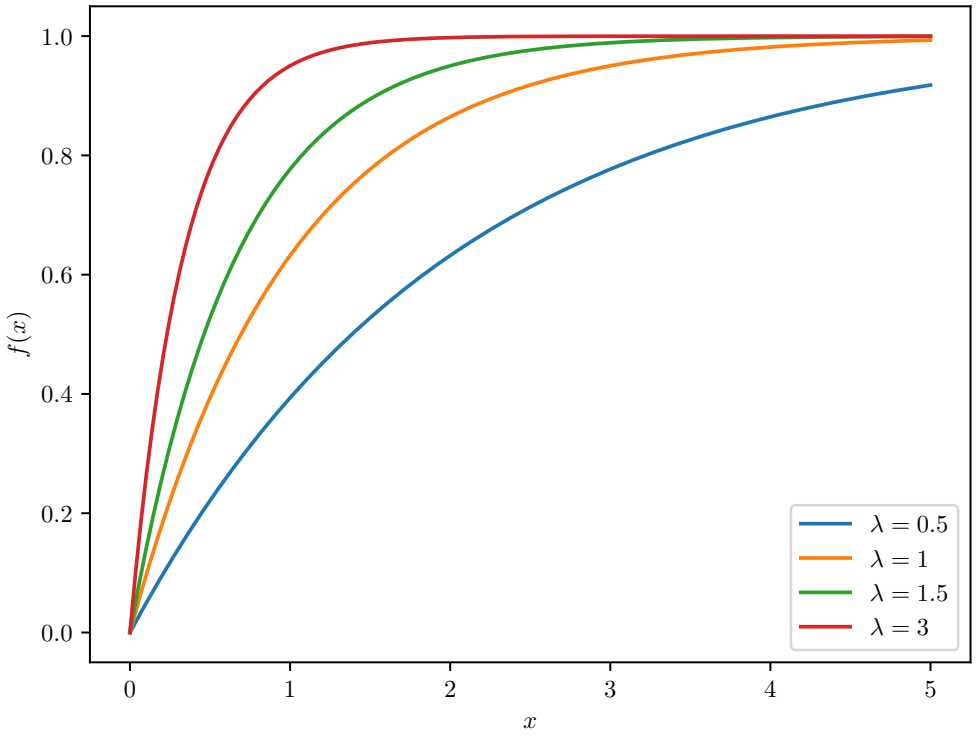


Figure 7.2: $F(x;\lambda) = 1 - e^{-\lambda x}$ if $x \geq 0$

7.1.3 Memorylessness

See https://en.wikipedia.org/wiki/Exponential_distribution

7.2 Moments

Mean	$\frac{1}{\lambda}$
Variance	$\frac{1}{\lambda^2}$

Chapter 8

Gamma distribution

8.1 Description

The gamma distribution was created to predict the wait time until future events. While the exponential distribution predicts the wait time until the **first** event happens, the gamma distribution predicts the wait time until the ***k*th** event occurs.

TODO: https://towardsdatascience.com/gamma-distribution-intuition-derivation-and-examples-55f407423840

We represent the gamma distribution by $\gamma(p, \alpha)$.

8.1.1 Probability density function

For $\alpha > 0$ and $p > 0$:

$$f(x) = \frac{\alpha^p}{\Gamma(p)} x^{p-1} e^{-\alpha x}, \text{ for } x > 0$$

Note that $f(x) > 0$ for $x > 0$, and:

$$\int_0^\infty f(x)\, dx = \frac{1}{\Gamma(p)} \int_0^\infty y^{p-1} e^{-y}\, dy = 1$$

because the definition of $\Gamma(p)$ is the value of the last integral.

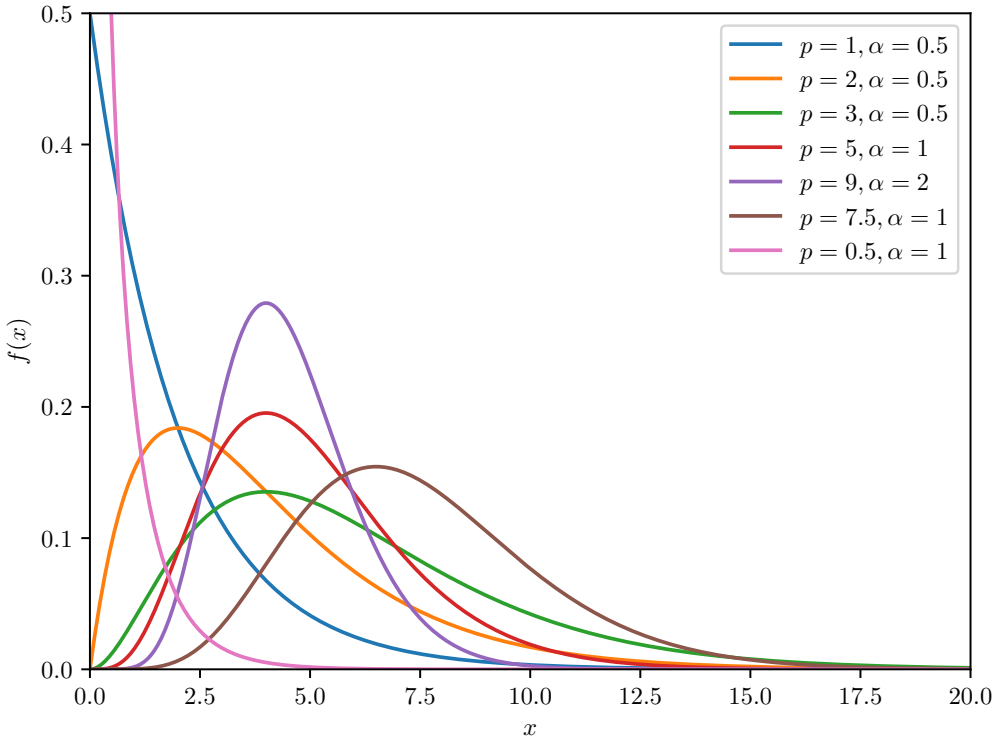


Figure 8.1: $f(x) = \frac{\alpha^p}{\Gamma(p)} x^{p-1} e^{-\alpha x}$, for $x > 0$, pdf of $\gamma(p, \alpha)$

8.1.2 Cumulative distribution function

For $\alpha > 0$ and $p > 0$:

$$F(x) = \int_0^x \frac{\alpha^p}{\Gamma(p)} t^{p-1} e^{-\alpha t}\, dt, \text{ for } x > 0$$

F can't be expressed in terms of elementary functions, but its values can be calculated.

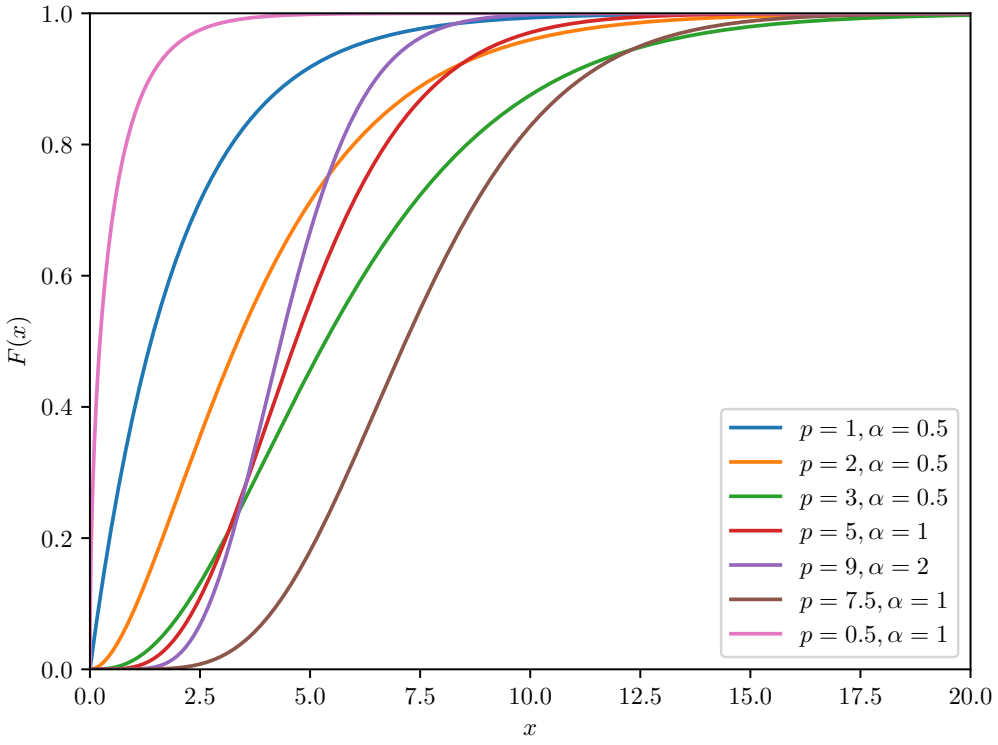


Figure 8.2: $F(x) = \int_0^x \frac{\alpha^p}{\Gamma(p)} t^{p-1} e^{-\alpha t}\, dt$, for $x > 0$, cdf of $\gamma(p, \alpha)$

8.2 Properties

- For $p = 1$, the Gamma distribution is the same as the exponential one.
- For $p < 1$, the density of Gamma distribution is unbounded when $x \to 0^+$, so the formula is an improper integral (which converges).
- For $p > 1$, the density function tends to 0 as $x \to 0^+$.

The parameter p determines the shape of the density function. The parameter α, however, is a scale parameter which expands or contracts the x axis. Indeed, if we use the change of variables $y = x/a$, with $a > 0$, we see that:

$$\frac{\alpha^p}{\Gamma(p)} x^{p-1} e^{-\alpha x}\, dx \quad \text{turns into} \quad \frac{(a\alpha)^p}{\Gamma(p)} y^{p-1} e^{-\alpha a y}\, dy$$

so that the probability that $\gamma(p, \alpha)$ assigns to the neighborhood of a point x is the same that $\gamma(p, \alpha a)$ assigns to the neighborhood of the point $y = x/a$. In particular,

if $a = 1/\alpha$, the distribution $\gamma(p, \alpha)$ becomes $\gamma(p, 1)$ with the change of variables $y = \alpha x$.

8.3 Moments

Mean	??
Variance	??

Part IV

Appendices

Appendix A

The Gamma and Beta functions

A.1 The Gamma function

A.1.1 Definition

The *gamma function* was introduced by Leonhard Euler in his goal to generalize the factorial to non-integer values.

Definition 40 (Euler, 1970)**.** Let $x > 0$. We define:

$$\Gamma(x) = \int_0^1 (-\log t)^{x-1}\, dt$$

Theorem 41. *For $x > 0$,*

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$$

Proof. Use the change of variable $u = -\log t,\ t = e^{-u},\ dt = -e^{-u}du$. Note that $t \to 0 \implies u \to \infty$, and $t \to 1 \implies u \to 0$. □

The derivatives can be deduced by differentiating under the integral sign of the second expression:

$$\Gamma'(x) = \int_0^\infty t^{x-1} e^{-t} \log(t)\, dt$$

$$\Gamma^{(n)}(x) = \int_0^\infty t^{x-1} e^{-t} \log^n(t)\, dt$$

Remark 42. Remember that $t = e^{\log(t)}$, and then: $t^{x-1} = (e^{\log(t)})^{x-1} = e^{\log(t)(x-1)}$. Thus, $\frac{d}{dx}t^{x-1} = \log(t)e^{\log(t)(x-1)} = \log(t)t^{x-1}$

A.1.2 Properties

The functional equation

We have that

$$\Gamma(1) = \int_0^\infty e^{-t}\,dt = 1$$

and, for $x > 0$, an integration by parts ($u = t^{x-1}, \frac{du}{dt} = t^{x-2}, v = e^{-t}, \frac{dv}{dt} = -e^{-t}$) yields:

$$\Gamma(x+1) = \int_0^\infty t^x e^{-t}\,dt = [-t^x e^{-t}]_0^\infty + x\int tx - 1e^{-t}\,dt = x\Gamma(x)$$

and the relation $\Gamma(x+1) = x\Gamma(x)$ is the important **functional equation**.

For integers the functional equation becomes:

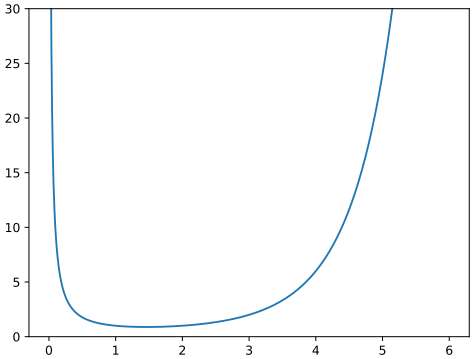
$$\Gamma(n+1) = n!$$

and it's why the gamma function can be seen as an extension of the factorial function to real non null positive numbers.

Other important values

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty \frac{e^{-t}}{\sqrt{t}}\,dt = 2\int_0^\infty e^{-u^2}\,du = 2\frac{\sqrt{\pi}}{2} = \sqrt{\pi}$$

Plot



A.2 The beta function

A.2.1 Definition

We define the *beta function* as

$$\beta(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$$

A.2.2 Properties

Proposition 43. *The beta function is symmetric:*

$$\beta(x,y) = \beta(y,x)$$

Proof. Because of the convergent property of definite integrals,

$$\int_0^a f(t)\,dt = \int_0^a f(a-t)\,dt$$

we can rewrite the above integral as

$$\beta(x,y) = \int_0^1 (1-t)^{x-1}t^{y-1}\,dt = \int_0^1 t^{y-1}(1-t)^{x-1}\,dt = \beta(y,x)$$

□

Proposition 44. *The following relation between the gamma and beta functions holds:*

$$\beta(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

Then, for positive integers x and y we can define the beta function as:

$$\beta(x,y) = \frac{(x-1)!(y-1)!}{(x+y-1)!}$$

Proof. Using the definition of the gamma function, we can write

$$\Gamma(m)\Gamma(n) = \int_0^\infty x^{m-1}e^{-x}\,dx \int_0^\infty y^{n-1}e^{-y}\,dy$$

Then we can rewrite it as a double integral

$$\Gamma(m)\Gamma(n) = \int_0^\infty \int_0^\infty x^{m-1}y^{n-1}e^{-(x+y)}\,dxdy$$

Applying the substitution $x = vt$ and $y = v(1-t)$, we have ... □

https://brilliant.org/wiki/beta-function/

Bibliography

[1] Ricardo Vélez Ibarrola. *Cálculo de probabilidades 2*. Ediasa, 2004.

[2] *Introduction to the Gamma Function*. National Taiwan University. 2002. URL: https://www.csie.ntu.edu.tw/~b89089/link/gammaFunction.pdf (visited on 07/18/2020).

[3] Aerin Kim. *List of Aerin Kim statistics and probability articles*. 2020. URL: https://towardsdatascience.com/@aerinykim (visited on 07/24/2020).