6. Distribution and Quantile Functions

As usual, our starting point is a random experiment with probability measure $\mathbb{P}$ on an underlying sample space. In this section, we will study two types of functions that can be used to specify the distribution of a random variable.

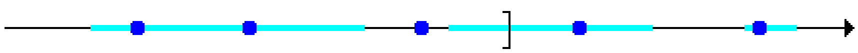
Distribution Functions

Definitions

Suppose that X is a real-valued random variable. The (cumulative) **distribution function** of X is the function F given by

$$F(x) = \mathbb{P}(X \le x), \quad x \in \mathbb{R}$$

This function is important because it makes sense for any type of random variable, regardless of whether the distribution is discrete, continuous, or even mixed, and because it completely determines the distribution of X. In the picture below, the light shading is intended to represent a continuous distribution of probability, while the darker dots represents points of positive probability; $F(x)$ is the total probability mass to the left of (and including) x.



We will abbreviate some limits of F as follows:

- $F(x^+) = \lim_{t \downarrow x} F(t)$
- $F(x^-) = \lim_{t \uparrow x} F(t)$
- $F(\infty) = \lim_{t \uparrow \infty} F(t)$
- $F(-\infty) = \lim_{t \downarrow -\infty} F(t)$

Basic Properties

The properties in the following exercise completely characterize distribution functions.

1. Show that F is increasing: if $x \le y$ then $F(x) \le F(y)$.

2. Show that $F(x^+) = F(x)$ for $x \in \mathbb{R}$. Thus, F is **continuous from the right**.

   a. Fix $x \in \mathbb{R}$. Let $x_1 > x_2 > \cdots$ be a decreasing sequence with $x_n \to x$ as $n \to \infty$.
   b. Show that the intervals $(-\infty, x_n]$ are decreasing in n and have intersection $(-\infty, x]$.
   c. Use the continuity theorem for decreasing events.

3. Show that $F(x^-) = \mathbb{P}(X < x)$ for $x \in \mathbb{R}$. Thus, F has **limits from the left**.

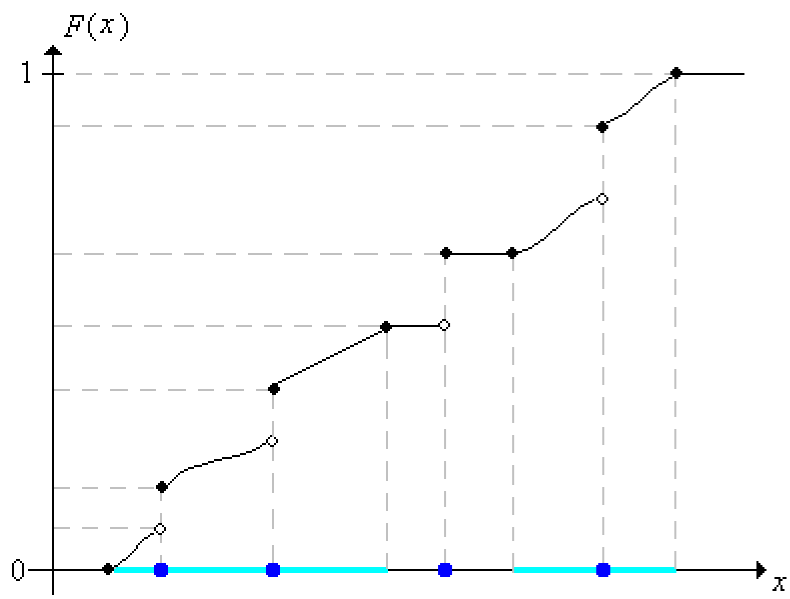
a. Fix $x \in \mathbb{R}$. Let $x_1 < x_2 < \cdots$ be an increasing sequence with $x_n \to x$ as $n \to \infty$.
b. Show that the intervals $(-\infty, x_n]$ are increasing in n and have union $(-\infty, x)$.
c. Use the continuity theorem for increasing events.

4. Show that $F(-\infty) = 0$.

a. Let $x_1 > x_2 > \cdots$ be a decreasing sequence with $x_n \to -\infty$ as $n \to \infty$.
b. Show that the intervals $(-\infty, x_n]$ are decreasing in n and have intersection $\varnothing$.
c. Use the continuity theorem for decreasing events.

5. Show that $F(\infty) = 1$.

a. Let $x_1 < x_2 < \cdots$ be an increasing sequence with $x_n \to \infty$ as $n \to \infty$.
b. Show that the intervals $(-\infty, x_n]$ are increasing in n and have union $\mathbb{R}$.
c. Use the continuity theorem for increasing events.



The following exercise shows how the distribution function can be used to compute the probability that X is in an interval. Recall that a probability distribution on $\mathbb{R}$ is completely determined by the probabilities of *intervals*; thus, the *distribution function* determines the *distribution* of X. In each of case, the main tool that you need is the difference rule:

$$\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A) \text{ for } A \subseteq B$$

6. Suppose that $a < b$. Show that

a. $\mathbb{P}(X = a) = F(a) - F(a^-)$
b. $\mathbb{P}(a < X \leq b) = F(b) - F(a)$

c. $\mathbb{P}(a < X < b) = F(b^-) - F(a)$
d. $\mathbb{P}(a \le X \le b) = F(b) - F(a^-)$
e. $\mathbb{P}(a \le X < b) = F(b^-) - F(a^-)$

Conversely, if a Function F on $\mathbb{R}$ satisfies the properties in Exercises 1-5, then the formulas in Exercise 6 define a probability distribution on $\mathbb{R}$, with F as the distribution function.

7. Show that if X has a continuous distribution, then the distribution function F is continuous in the usual calculus sense. Thus, the two meanings of *continuous* come together.

Relation to Density Functions

There are simple relationships between the distribution function and the probability density function.

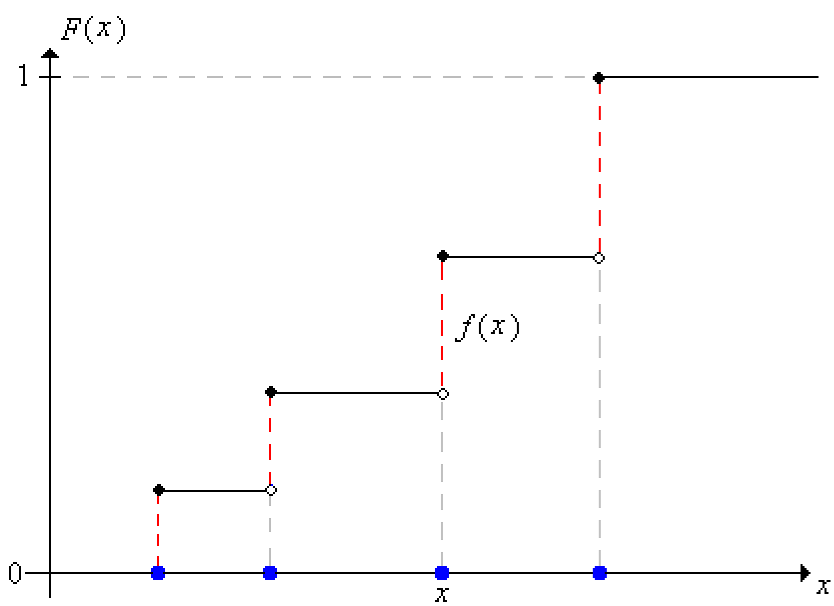
8. Suppose that X has discrete distribution on a countable subset $S \subseteq \mathbb{R}$. Let f denote the probability density function and F the distribution function. Show that for $x \in \mathbb{R}$,

$$F(x) = \sum_{(t \in S) \text{ and } (t \le x)} f(t)$$

Conversely, show that for $x \in S$,

$$f(x) = F(x) - F(x^-)$$

Thus, F is a step function with jumps at the points in S; the size of the jump at x is the value of the probability density function at x.



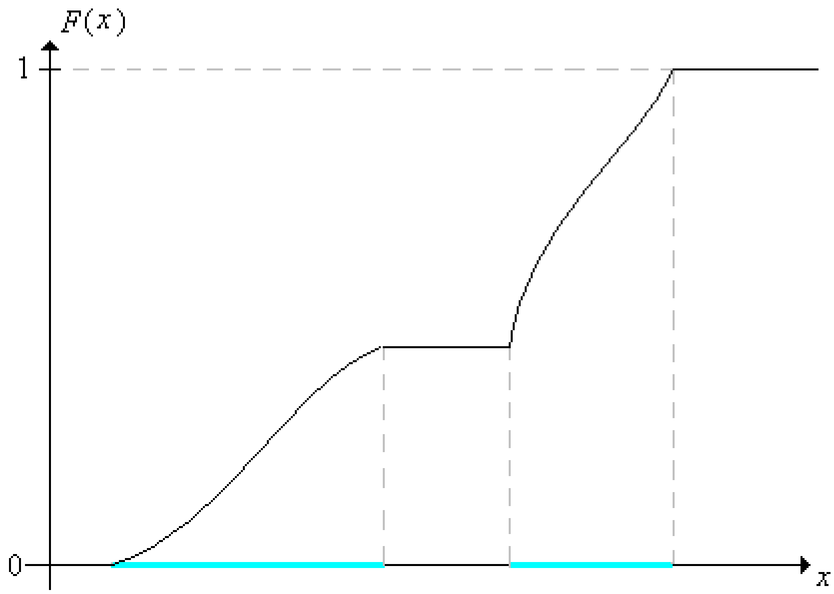
There is an analogous result for a continuous distribution with a probability density function.

9. Suppose that X has a continuous distribution on $\mathbb{R}$ with probability density function f (which we will assume is

piecewise continuous) and distribution function F. Show that for $x \in \mathbb{R}$,

$$F(x) = \int_{-\infty}^{x} f(t) dt$$

Conversely, show that if f is continuous at x, then F is differentiable at x and $f(x) = F'(x)$



The result in the last exercise is the basic probabilistic version of the fundamental theorem of calculus. For mixed distributions, we have a combination of the results in the last two exercises.

10. Suppose that X has a mixed distribution, with discrete part on a countable subset $S \subseteq \mathbb{R}$, and continuous part on $\mathbb{R}$. Let g denote the partial probability density function of the discrete part, h the partial probability density function of the continuous part, and F the distribution function. Show that

a. $F(x) = \sum_{(t \in S) \text{ and } (t \le x)} g(t) + \int_{-\infty}^{x} h(t) dt$ for $x \in \mathbb{R}$,
b. $g(x) = F(x) - F(x^-)$ for $x \in S$
c. $h(x) = F'(x)$ if $x \notin S$ and h is continuous at x,

Naturally, the distribution function can be defined relative to any of the conditional distributions we have discussed. No new concepts are involved, and all of the results above hold.

11. Suppose that X has a continuous distribution on $\mathbb{R}$ with density function f that is symmetric about a point a:

$$f(a + t) = f(a - t), \quad t \in \mathbb{R}$$

Show that the distribution function F satisfies

$$F(a - t) = 1 - F(a + t), \quad t \in \mathbb{R}$$

Reliability

Suppose again that X is a random variable with distribution function F. A function that clearly gives the same information as F is the **right tail distribution function**:

$$G(x) = 1 - F(x) = \mathbb{P}(X > x), \quad x \in \mathbb{R}$$

.

12. Give the mathematical properties of a right tail distribution function, analogous to the properties in Exercise 1.

Suppose that T is a random variable with a continuous distribution on $[0, \infty)$. If we interpret T as the lifetime of a device, then the right tail distribution function G is called the **reliability function**: $G(t)$ is the probability that the device lasts at least t time units. Moreover, the function h defined below is called the **failure rate function**:

$$h(t) = \frac{f(t)}{G(t)}, \quad t > 0$$

13. Show that $h(t)\,dt \approx \mathbb{P}(t < T < t + dt | T > t)$ if dt is small.

Thus, $h(t)\,dt$ is the probability that the device will fail in next dt time units, given survival up to time t. Moreover, the failure rate function completely determines the distribution of T.

14. Show that

$$G(t) = \exp\left(-\int_0^t h(s) ds\right), \quad t > 0$$

15. Show that the failure rate function h satisfies the following properties:

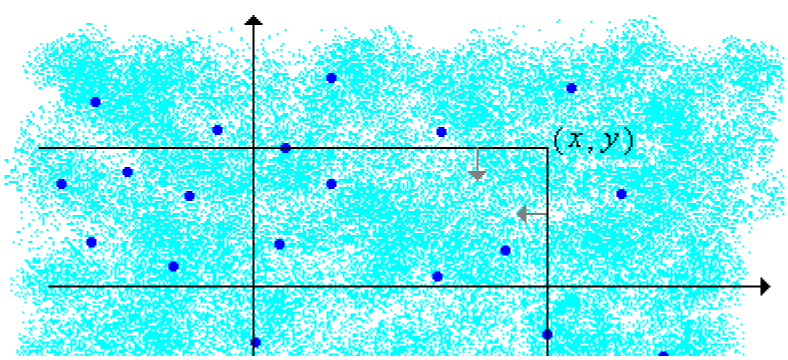
a. $h(t) \geq 0, \; t > 0$
b. $\int_0^\infty h(t) dt = \infty$

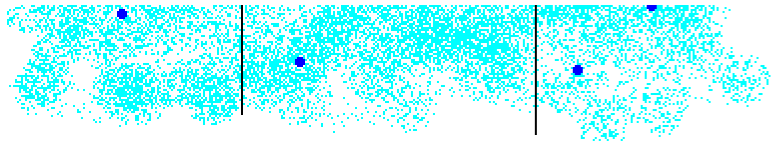
16. Conversely, suppose that h satisfies the conditions in Exercise 15. Show that the formula in Exercise 14 defines a reliability function.

Multivariate Distribution Functions

Suppose that X and Y are real-valued random variables for an experiment, so that (X, Y) is random vector taking values in a subset of $\mathbb{R}^2$. The **distribution function** of (X, Y) is the function F defined by

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y), \; (x, y) \in \mathbb{R}^2$$





In the graph above, the light shading is intended to suggest a continuous distribution of probability, while the darker dots represent points of positive probability. Thus, $F(x, y)$ is the total probability mass below and to the left (that is, southwest) of the point (x, y). As in the single variable case, the distribution function of (X, Y) completely determines the distribution of (X, Y).

17. Suppose that a, b, c, d are real numbers with $a < b$ and $c < d$. Show that

$$\mathbb{P}(a < X \le b, c < Y \le d) = F(b, d) - F(a, d) - F(b, c) + F(a, c)$$

18. In the setting of the previous exercise, give the generalizations of the results in Exercise 6. Than is, give the appropriate formula on the right for all possible combinations of weak and strong inequalities on the left.

19. Let F denote the distribution function of (X, Y) let G and H denote the distribution functions of X and Y, respectively. Show that

a. $G(x) = F(x, \infty)$
b. $H(y) = F(\infty, x)$

20. In the context of the last exercise, show that X and Y are independent if and only if

$$F(x, y) = G(x) H(y) \text{ for all } (x, y) \in \mathbb{R}^2$$

All of the results of this subsection generalize in a straightforward way to n-dimensional random vectors.

The Empirical Distribution Function

Suppose that $\{x_1, x_2, ..., x_n\}$ is a data set of observed values from a real-valued random variable. The **empirical distribution function** is defined by

$$F_n(x) = \frac{\#(\{i \in \{1, 2, ..., n\} : x_i \le x\})}{n}, \quad x \in \mathbb{R}$$

Thus, $F_n(x)$ gives the proportion of values in the data set that are less than or equal to x.

Quantile Functions

Definitions

Let X be a random variable with distribution function F, and let $p \in (0, 1)$. A value of x such that

$$F(x^-) = \mathbb{P}(X < x) \le p \text{ and } F(x) = \mathbb{P}(X \le x) \ge p$$

is called a **quantile** of order p for the distribution. Roughly speaking, a quantile of order p is a value where the graph of the cumulative distribution function crosses (or jumps over) p. For example, in the picture below, a is the unique quantile of order p and b is the unique quantile of order q. On the other hand, the quantiles of order r form the interval $[c, d]$, and moreover, d is a quantile for all orders in the interval $[r, s]$.

Note that there is an inverse relation of sorts between the quantiles and the cumulative distribution values, but the relation is more complicated than that of a function and its ordinary inverse function, because the distribution function is not one-to-one in general. For many purposes, it is helpful to select a specific quantile for each order; to do this requires defining a **generalized inverse** of the distribution function.

21. Let $p \in (0, 1)$. Use the fact that F is right continuous and increasing to show that $\{x \in \mathbb{R} : F(x) \ge p\}$ is an interval of the form $[a, \infty)$.

Thus, we define the **quantile function**; by

$$F^{-1}\left(p\right) = \min\{x \in \mathbb{R} : F(x) \ge p\}, \quad p \in (0, 1)$$

Note that the definition makes sense by Exercise 21. Note also that if F strictly increases from 0 to 1 on an interval S (so that the underlying distribution is continuous and is supported on S), then F^{-1} is the ordinary inverse of F. We do not usually define the quantile function at the endpoints 0 and 1. If we did, note that $F^{-1}\left(0\right)$ would always be $-\infty$.

Properties

The following exercise justifies the name: $F^{-1}\left(p\right)$ is the minimum of the quantiles of order p.

22. For $p \in (0, 1)$ show that

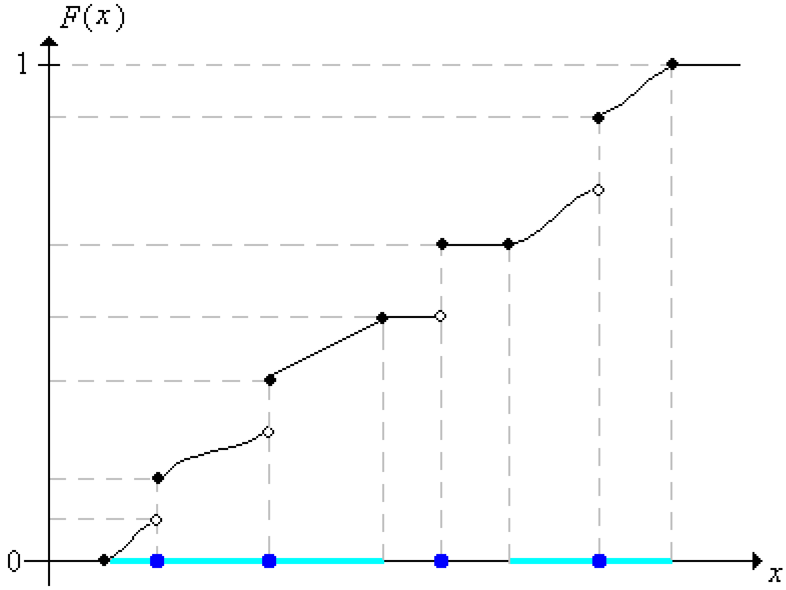
a. $F^{-1}\left(p\right)$ is a quantile of order p.
b. If x is a quantile of order p then $F^{-1}\left(p\right) \le x$

The other basic properties are given in the following two exercises.

23. Show that

a. F^{-1} is increasing on $(0, 1)$.
b. $F^{-1}\left(F(x)\right) \le x$ for any $x \in \mathbb{R}$ with $0 < F(x) < 1$.
c. $F\left(F^{-1}\left(p\right)\right) \ge p$ for any $p \in (0, 1)$
d. $F^{-1}\left(p^-\right) = F^{-1}\left(p\right)$ for $p \in (0, 1)$. Thus F^{-1} is continuous from the left.
e. $F^{-1}\left(p^+\right) = \inf\{x \in \mathbb{R} : F(x) > p\}$ for $p \in (0, 1)$. Thus F^{-1} is has limits from the right.

As always, the inverse of a function is obtained essentially by reversing the roles of independent and dependent variables. In the graphs below, note that jumps of F become flat portions of F^{-1} while flat portions of F become jumps of F^{-1}



24. Show that for $x \in \mathbb{R}$ and $p \in (0, 1)$, $F^{-1}(p) \le x$ if and only if $p \le F(x)$

Special Quantiles

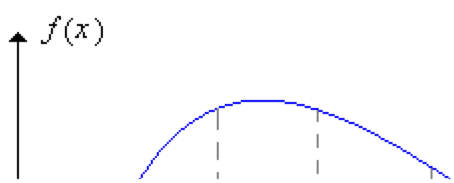
A quantile of order $\frac{1}{2}$ is called a **median** of the distribution. When there is only one median, it is frequently used as a measure of the **center** of the distribution. A quantile of order $\frac{1}{4}$ is called a **first quartile** and the quantile of order $\frac{3}{4}$ is called a **third quartile**. A median is a second quartile. Assuming uniqueness, let q_1, q_2, and q_3 denote the first, second, and third quartiles of X, respectively. Note that the interval from q_1 to q_3, gives the middle half of the distribution, and thus the **interquartile range** is defined to be

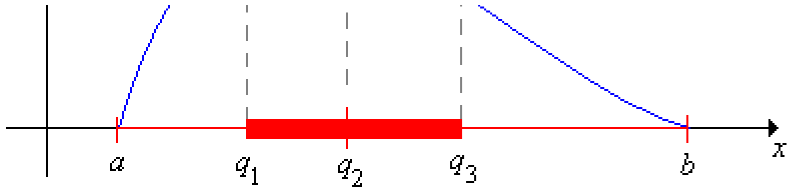
$$\text{IQR} = q_3 - q_1$$

and is sometimes used as a measure of the **spread** of the distribution with respect to the median. Let a and b denote the minimum and maximum values of X, respectively (assuming that these are finite). The five parameters

$$(a, q_1, q_2, q_3, b)$$

are often referred to as the **five-number summary**. Together, these parameters give a great deal of information about the distribution in terms of the center, spread, and skewness. Graphically, the five numbers are often displayed as a **boxplot**, which consists of a line extending from the minimum value a to the maximum value b, with a rectangular box from q_1 to q_3, and tick marks at a, the median q_2, and b. Roughly speaking, the five numbers separate the set of values of X into 4 intervals of approximate probability $\frac{1}{4}$ each.





25. Suppose that X has a continuous distribution on $\mathbb{R}$ with density f that is symmetric about a point a, that is $f(a - t) = f(a + t)$, $t \in \mathbb{R}$. Show that if $a + t$ is a quantile of order p then $a - t$ is a quantile of order $1 - p$.

Examples and Applications

26. Let F be the function defined by

$$F(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{10}, & 1 \le x < \frac{3}{2} \\ \frac{3}{10}, & \frac{3}{2} \le x < 2 \\ \frac{6}{10}, & 2 \le x < \frac{5}{2} \\ \frac{9}{10}, & \frac{5}{2} \le x < 3 \\ 1, & x \ge 3 \end{cases}$$

a. Sketch the graph of F and show that F is the distribution function for a discrete distribution.
b. Find the corresponding density function f and sketch the graph.
c. Find $\mathbb{P}(2 \le X < 3)$ where X has this distribution.
d. Find the quantile function and sketch the graph.
e. Find the five number summary and sketch the boxplot.

27. Let F be the function defined by

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{x}{x + 1}, & x \ge 0 \end{cases}$$

a. Sketch the graph of F and show that F is the distribution function for a continuous distribution.
b. Find the corresponding density function f and sketch the graph.
c. Find $\mathbb{P}(2 \le X < 3)$ where X has this distribution.
d. Find the quantile function and sketch the graph.
e. Find the five number summary and sketch the boxplot.

28. Let F be the function defined by

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{4}x, & 0 \le x < 1 \\ \frac{1}{3} + \frac{1}{4}(x-1)^2, & 1 \le x < 2 \\ \frac{2}{3} + \frac{1}{4}(x-2)^3, & 2 \le x < 3 \\ 1, & x \ge 3 \end{cases}$$

a. Sketch the graph of F and show that F is the distribution function of a mixed distribution.
b. Find the partial density of the discrete part and sketch the graph.
c. Find the partial density of the continuous part and sketch the graph.
d. Find $\mathbb{P}(2 \le X < 3)$ where X has this distribution.
e. Find the quantile function and sketch the graph.
f. Find the five number summary and sketch the boxplot.

The Uniform Distribution

29. Suppose that X has probability density function $f(x) = \frac{1}{b-a}$, $a \le x \le b$. Thus, X is uniformly distributed on the interval $[a, b]$.

a. Find the distribution function and sketch the graph.
b. Find the quantile function and sketch the graph.
c. Compute the five-number summary.
d. Sketch the graph of the probability density function with the boxplot on the horizontal axis.

The Exponential Distribution

30. Suppose that T has probability density function $f(t) = r\,e^{-r\,t}$, $t \ge 0$, where $r > 0$ is a parameter.

a. Find the distribution function and sketch the graph.
b. Find the reliability function and sketch the graph.
c. Find the failure rate function and sketch the graph.
d. Find the quantile function and sketch the graph.
e. Compute the five-number summary.
f. Sketch the graph of the probability density function with the boxplot on the horizontal axis.

The distribution in the last exercise is the **exponential distribution** with rate parameter r. Note that this distribution is characterized by the fact that it has **constant failure rate**. The exponential distribution is studied in detail in the chapter on The Poisson Process.

31. In the quantile applet, select the gamma distribution and set the shape parameter $k = 1$ to get the exponential distribution. Vary the scale parameter b (the reciprocal of the rate parameter) and note the shape of the density function and the distribution function.

The Pareto Distribution

32. Suppose that X has probability density function $f(x) = \frac{a}{x^{a+1}}$ for $x \geq 1$ where $a > 0$ is a parameter.

a. Find the distribution function.
b. Find the reliability function.
c. Find the failure rate function.
d. Find the quantile function.
e. Compute the five-number summary.
f. In the case $a = 2$, sketch the graph of the probability density function with the boxplot on the horizontal axis.

The distribution in the last exercise is the **Pareto distribution** with shape parameter a, named after Vilfredo Pareto. The Pareto distribution is studied in detail in the chapter on Special Distributions.

33. In the quantile applet, select the Pareto distribution. Vary the parameter and note the shape of the density function and the distribution function.

The Cauchy Distribution

34. Suppose that X has density function $f(x) = \frac{1}{\pi\left(1+x^2\right)}, \quad x \in \mathbb{R}$.

a. Find the distribution function and sketch the graph.
b. Find the quantile function and sketch the graph.
c. Compute the five-number summary and the interquartile range.
d. Sketch the graph of the probability density function with the boxplot on the horizontal axis.

The distribution in the last exercise is the **Cauchy distribution**, named after Augustin Cauchy. The Cauchy distribution is a member of the **Student *t*** family of distributions. The Student *t* distributions are studied in detail in the chapter on Special Distributions.

35. In the quantile applet, select the student distribution and set the degrees of freedom parameter to 1 to get the Cauchy distribution. Note the shape of the density function and the distribution function.

The Weibull Distribution

36. Let $h(t) = k\, t^{k-1}$, $t > 0$ where $k > 0$ is a parameter.

a. Sketch the graph of h in the three cases $0 < k < 1$, $k = 1$, $k > 1$.
b. Show that h is a failure rate function.
c. Find the reliability function and sketch the graph in each of the three cases.
d. Find the distribution function and sketch the graph in each of the three cases.
e. Find the density function and sketch the graph in each of the three cases.
f. Find the quantile function and sketch the graph in each of the three cases.
g. Compute the five-number summary.

The distribution in the previous exercise is the **Weibull distributions** with shape parameter k, named after Walodi Weibull. The Weibull distribution is studied in detail in the chapter on Special Distributions.

37. In the quantile applet, select the Weibull distribution. Vary the parameters and note the shape of the density function and the distribution function.

Beta Distributions

38. Suppose that X has density function $f(x) = 12\, x^2\, (1 - x)$, $0 \le x \le 1$.

a. Find the distribution function of X and sketch the graph.
b. Compute the five number summary and the interquartile range. You may have to approximate the quantiles.
c. Sketch the graph of the density function with the boxplot on the horizontal axis.

39. Suppose that X has density function $f(x) = \dfrac{1}{\pi\sqrt{x\,(1-x)}}$, $0 < x < 1$.

a. Find the distribution function of X and sketch the graph.
b. Find the quantile function and sketch the graph.
c. Compute the five number summary and the interquartile range.
d. Sketch the graph of the density function with the boxplot on the horizontal axis.

The distributions in the last two exercises are examples of **beta distributions**. The particular beta distribution in the last exercise is also known as the **arcsine distribution**; the distribution function explains the name. Beta distributions are studied in detail in the chapter on Special Distributions.

40. In the quantile applet, select the beta distribution. For each of the following parameter values, note the location and shape of the density function and the distribution function.

a. $a = 3$, $b = 2$. This gives the beta distribution in Exercise 38.
b. $a = \frac{1}{2}$, $b = \frac{1}{2}$. This gives the arcsine distribution in Exercise 39.

Logistic Distribution

41. Let $F(x) = \frac{e^x}{1+e^x}$, $x \in \mathbb{R}$.

a. Show that F is a distribution function for a continuous distribution, and sketch the graph.
b. Find the quantile function and sketch the graph.
c. Compute the five-number summary and the interquartile range.
d. Find the density function and sketch the graph with the boxplot on the horizontal axis.

The distribution in the last exercise is an **logistic distribution** and the quantile function is known as the **logit function**. The logistic distribution is studied in detail in the chapter on Special Distributions.

42. In the quantile applet, select the logistic distribution and note the shape of the density function and the distribution function.

Extreme Value Distribution

43. Let $F(x) = e^{-e^{-x}}$, $x \in \mathbb{R}$.

a. Show that F is a distribution function for a continuous distribution, and sketch the graph.
b. Find the quantile function and sketch the graph.
c. Compute the five-number summary.
d. Find the density function and sketch the graph with the boxplot on the horizontal axis.

The distribution in the last exercise is the **type 1 extreme value distribution**, also known as the **Gumbel distribution** in honor of Emil Gumbel. Extreme value distributions are studied in detail in the chapter on Special Distributions.

44. In the quantile applet, select the extreme value distribution and note the shape and location of the density function and the distribution function.

---

45. Suppose that X has probability density function $f(x) = -\ln(x)$, $0 < x < 1$.

a. Sketch the graph of f
b. Find the distribution function F and sketch the graph.

c. Find $\mathbb{P}\left(\frac{1}{3} \le X \le \frac{1}{2}\right)$

46. In the histogram applet, select boxplot. Set the class width to 0.1 and construct a discrete distribution with at least 30 values of each of the types indicated below. Note the shape of the boxplot and the relative positions of the parameters in the five-number summary:

a. A uniform distribution.
b. A symmetric, unimodal distribution.
c. A unimodal distribution that is skewed right.
d. A unimodal distribution that is skewed left.
e. A symmetric bimodal distribution.
f. A u-shaped distribution.

47. Suppose that a pair of fair dice are rolled and the sequence of scores (X_1, X_2) is recorded.

a. Find the distribution function of $Y = X_1 + X_2$, the sum of the scores.
b. Find the distribution function of $V = \max\{X_1, X_2\}$, the maximum score.
c. Find the conditional distribution function of Y given $V = 5$

48. Suppose that (X, Y) has density function $f(x, y) = x + y, \ 0 < x < 1, \ 0 < y < 1$.

a. Find the distribution function of (X, Y).
b. Find the distribution function of X.
c. Find the distribution function of Y.
d. Find the conditional distribution function of X given $Y = y$ for $0 < y < 1$.
e. Find the conditional distribution function of Y given $X = x$ for $0 < x < 1$.
f. Are X, Y independent?

Statistical Exercises

49. For the M&M data, compute the empirical distribution function of the total number of candies.

50. For the cicada data, let L denotes body length and let G denote gender. Compute the empirical distribution function of the following variables:

a. L
b. L given G = male
c. L given G = female.
d. Do you believe that L and G are independent?

For statistical versions of some of the topics in this section, see the chapter on Random Samples, and in particular, the sections on empirical distributions and order statistics.