

STAT 509: STATISTICS FOR ENGINEERS

Fall 2025

**Lecture Notes
Sections 001 and 007**

**Joshua M. Tebbs
Department of Statistics
University of South Carolina**

© by Joshua M. Tebbs

Contents

1	Introduction	1
2	Probability	5
2.1	Sample spaces and events	7
2.2	Counting techniques	9
2.3	Axioms of probability and additive rules	13
2.4	Conditional probability and independence	15
2.5	Law of Total Probability and Bayes' Rule	19
2.6	Introduction to random variables	21
3	Discrete Distributions	23
3.1	Probability mass functions	23
3.2	Mean and variance	26
3.3	Binomial distribution	33
3.4	Geometric and negative binomial distributions	38
3.5	Hypergeometric distribution	42
3.6	Poisson distribution	45
4	Continuous Distributions	48
4.1	Probability density functions	48
4.2	Mean, variance, and percentiles	56
4.3	Exponential distribution	61
4.4	Gamma distribution	66
4.5	Normal distribution	69
5	Reliability Analysis and Lifetime Distributions	73
5.1	Weibull distribution	73
5.2	Reliability functions	77
5.3	Fitting a Weibull distribution to data	80
5.4	Quantile-quantile plots	84

6	Bridge to Statistical Inference	88
6.1	Populations and samples (Parameters and statistics)	88
6.2	Point estimation and sampling distributions	91
6.3	Sampling distribution of \bar{X}	94
6.4	The t distribution	101
6.5	Normal quantile-quantile plots	103
7	One-Sample Inference	106
7.1	Confidence interval for a population mean μ	107
7.2	Confidence interval for a population variance σ^2	112
7.3	Confidence interval for a population proportion p	119
7.4	Sample size determination	125

1 Introduction

Definition: Statistics is the science of data; how to interpret and visualize data, analyze data, and design studies to collect data.

- Statistics is used in all disciplines; not just in engineering.
- “Statisticians get to play in everyone else’s back yard.” (John Tukey)

Examples:

1. In a reliability (time to event) study, engineers are interested in describing the time until failure for a jet engine fan blade.
2. In a genetics study involving patients with Alzheimer’s disease, researchers wish to identify genes that are differentially expressed (when compared to non-AD patients).
3. In an agricultural experiment, researchers want to know which of four fertilizers (which vary in their nitrogen levels) produces the highest corn yield.
4. In a clinical trial, physicians want to determine which of two drugs is more effective for managing weight loss in pre-diabetic patients.
5. In a public health study involving “at-risk” teenagers, epidemiologists want to know how smoking behavior differs across demographic classes.
6. A food scientist is interested in determining how different feeding schedules (for pigs) could affect the spread of salmonella during the slaughtering process.
7. A pharmacist wants to determine if administering caffeine to premature infants increases the risk of necrotizing enterocolitis.
8. A research dietician wants to determine if academic achievement is related to BMI among students in the fourth grade.

What we do: Statisticians (and data scientists) use their skills in mathematics and computing to formulate statistical models and analyze data for a specific problem at hand. These models are then used to estimate important quantities of interest, to test the validity of proposed conjectures, and to predict future behavior. Being able to identify and model sources of **variability** are critical parts of this process.

Definition: A **deterministic model** makes no attempt to explain variability.

- In chemistry, the ideal gas law states

$$PV = nRT,$$

where P = pressure of a gas, V = volume, n = amount of substance of gas (number of moles), R = Boltzmann’s constant, and T = temperature.

- In circuit analysis, Ohm's law states

$$V = IR,$$

where V = voltage, I = current, and R = resistance.

In both of these models, the relationship among the variables is completely determined without ambiguity. In real life, this is rarely true for the obvious reason: there is natural variation that arises in the measurement process.

- For example, a common electrical engineering experiment involves setting up a simple circuit with a known resistance R . For a given current I , different students (using a multimeter) will then measure the voltage V .
 - With a class of 20 students, conducting the experiment with the same current I , we might get 20 different voltage measurements V_1, V_2, \dots, V_{20} .
 - A deterministic model is too simplistic; it does not acknowledge the inherent variability that arises in the measurement process.

Important: Statistical models are not deterministic. They incorporate **variability** with the hope of describing what is going on in a larger population of individuals. They can also be used to **predict** future outcomes for specific individuals, a common task in machine learning and artificial intelligence.

Example 1.1. An article by Liu et al. (1996) in *Journal of Air and Waste Management Association* described a research study motivated by the waste disposal problems in Kaohsiung City, Taiwan. The goal was to develop a statistical model to explain how the response variable

Y = energy content of solid waste specimen when incinerated (kcal/kg)

was related to four independent variables measured on each waste specimen

- x_1 = plastic by weight (measured as % of total weight)
- x_2 = paper by weight (measured as % of total weight)
- x_3 = garbage by weight (measured as % of total weight)
- x_4 = moisture percentage.

One way to think about this modeling problem—from a purely mathematical point of view—is to assume there is a function f that links the response variable Y to the independent variables for all waste specimens that will ever be collected, say

$$Y = f(x_1, x_2, x_3, x_4).$$

This is how a mathematician might formulate the problem—by using a **deterministic model**. Of course, the overarching problem is that the function f is likely unknown in real life so hence the model is not all that helpful. This is where statistics comes in.

Table 1.1: Waste incineration data. Measurements of energy, plastic percentage, paper percentage, garbage percentage, and moisture for a sample of 30 waste specimens.

Specimen	Energy (Y)	Plastic (x_1)	Paper (x_2)	Garbage (x_3)	Moisture (x_4)
1	947	18.69	15.65	45.01	58.21
2	1407	19.43	23.51	39.69	46.31
3	1452	19.24	24.23	43.16	46.63
4	1553	22.64	22.20	35.76	45.85
5	989	16.54	23.56	41.20	55.14
6	1162	21.44	23.65	35.56	54.24
7	1466	19.53	24.45	40.18	47.20
8	1656	23.97	19.39	44.11	43.82
9	1254	21.45	23.84	35.41	51.01
10	1336	20.34	26.50	34.21	49.06
11	1097	17.03	23.46	32.45	53.23
12	1266	21.03	26.99	38.19	51.78
13	1401	20.49	19.87	41.35	46.69
14	1223	20.45	23.03	43.59	53.57
15	1216	18.81	22.62	42.20	52.98
16	1334	18.28	21.87	41.50	47.44
17	1155	21.41	20.47	41.20	54.68
18	1453	25.11	22.59	37.02	48.74
19	1278	21.04	26.27	38.66	53.22
20	1153	17.99	28.22	44.18	53.37
21	1225	18.73	29.39	34.77	51.06
22	1237	18.49	26.58	37.55	50.66
23	1327	22.08	24.88	37.07	50.72
24	1229	14.28	26.27	35.80	48.24
25	1205	17.74	23.61	37.36	49.92
26	1221	20.54	26.58	35.40	53.58
27	1138	18.25	13.77	51.32	51.38
28	1295	19.09	25.62	39.54	50.13
29	1391	21.25	20.63	40.72	48.67
30	1372	21.62	22.71	36.22	48.19

Discussion: A statistician will think of this problem as

$$Y = f(x_1, x_2, x_3, x_4) + \epsilon,$$

where the extra term ϵ incorporates all the sources of variability that make Y different than $f(x_1, x_2, x_3, x_4)$. This could include

- independent variables that are not accounted for in the research study
- measurement error (e.g., are all the percentages measured correctly? the energy content measurements?)

- natural sampling variability, that is, the variability that arises from looking at a sample of waste specimens taken from a larger population.
 - Different samples of waste specimens will produce different data sets!

All of these sources of variability make ϵ random. We call

$$Y = f(x_1, x_2, x_3, x_4) + \epsilon,$$

a **statistical (or probabilistic) model** to acknowledge this. Statistical models acknowledge the relationship between (or among) variables is not perfect. This is more realistic.

Q: So, how do we estimate f with the sample of waste specimens like those in Table 1.1?

A: We could estimate f nonparametrically. That is, make no assumption about the form of f and use the observed data to select the function f that most closely matches the data.

- This is a challenging problem statistically and would give rise to a “wiggly” highly nonlinear function of x_1, x_2, x_3 , and x_4 as the solution.
- The function chosen might be hard to interpret on practical grounds.
- Predictions might give far too much weight to the data. This would be bad if new waste specimens did not align with the sample.

A: Another approach (which is more common) is to make simplifying assumptions about the form of f , say, that f is linear function of the independent variables, that is,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon.$$

This is called a **multiple linear regression model**.

- The coefficients $\beta_0, \beta_1, \beta_2, \beta_3$, and β_4 are called **regression parameters**.
- They describe how Y is related to the independent variables in the population of all waste specimens.
- Therefore, under the simplifying assumption that the relationship between Y and the independent variables is linear, we have reduced the problem of “estimating f ” to finding values of $\beta_0, \beta_1, \beta_2, \beta_3$, and β_4 which “best fit” the observed data in Table 1.1. This problem can be solved easily using linear algebra calculations and will be discussed in Chapter 11.

Remark: Statistical models incorporate randomness. Randomness arises in probability, which is known as “the mathematics of uncertainty.” Therefore, probability and uncertainty form the basis for all statistical analyses.

- Chapters 2-5 deal with probability and probability models for single variables.
- Chapter 6 is a “bridge” chapter. Chapters 7-12 deal with statistical methods to analyze data.

2 Probability

Example 2.1. A local television station’s meteorologist announces,

“There is a 30 percent chance of rain tomorrow.”

How do you interpret this statement?

- (a) It will rain tomorrow for 30 percent of the time. That is, for 7.2 hours tomorrow, it will be raining. For the remaining 16.8 hours, it will not be raining.
- (b) It will rain tomorrow in 30 percent of the region covered by the local television station. It will not rain in the other 70 percent of the region.
- (c) Among all local meteorologists, 30 percent of them think it will rain tomorrow. The remaining 70 percent of the meteorologists think it will not rain tomorrow.
- (d) Thirty percent of all inhabitants of the region covered by this local television station will see rain at least once during their day tomorrow; the remaining 70 percent will not see rain during their day.
- (e) It will rain on 30 percent of the days in which this same forecast is made.

Discussion: The statement incorporates uncertainty about a future event—the event that it will rain tomorrow. The phrase “30 percent” can be interpreted as a probability. None of us know for sure whether it will rain tomorrow, so we are dealing with a **random** event. We can write this as

$$P(A) = 0.30,$$

where the event

$$A = \{\text{it rains tomorrow}\}.$$

In this example, we are given five different interpretations of $P(A) = 0.30$. I think interpretation (e) makes the most sense, but all five interpretations are valid depending on how one conceptualizes the problem.

Example 2.2. I have a class with 50 students in it. Assuming there are no twins (or other multiple births), what is the probability there is at least one shared birthday among the students? Remember there are 365 days in a year.

- (a) $P(A) = 0.14$
- (b) $P(A) = 0.28$
- (c) $P(A) = 0.45$
- (d) $P(A) = 0.81$
- (e) $P(A) = 0.97$

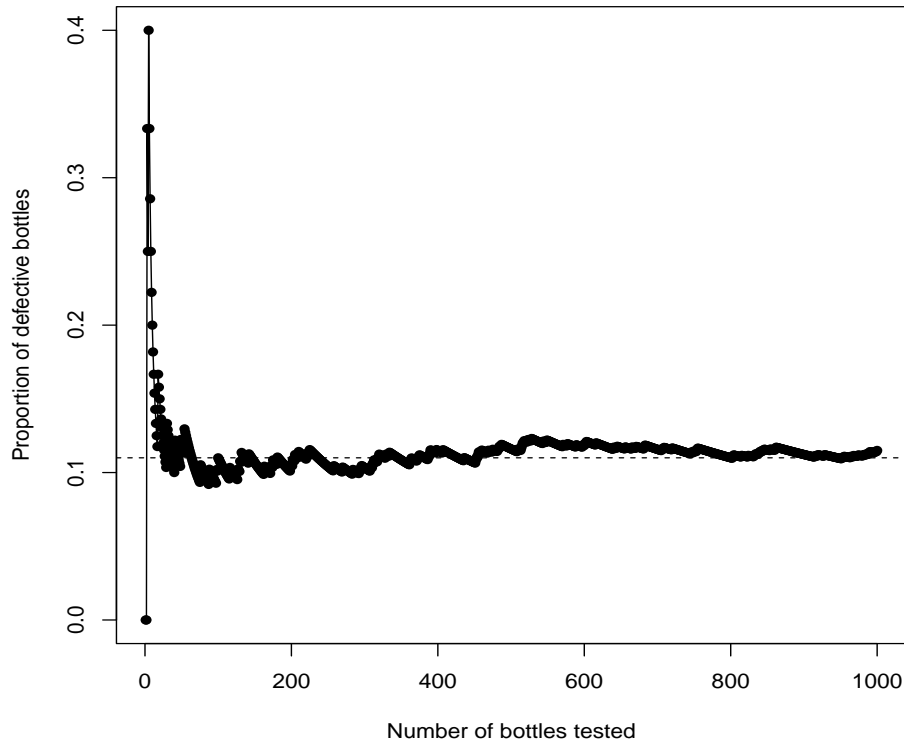


Figure 2.1: Relative frequency plot of the proportion of defective bottles (i.e., bottles that do not conform to specifications). A total of 1000 bottles are observed. A dotted horizontal line at 0.11 is added.

Q: For a future event A , how do we interpret what $P(A)$ means?

A: One way to think about $P(A)$ is that it represents the long-run proportion of times A would occur if we were to repeat the same phenomenon over and over again. The event A may occur on some repetitions and not on others; $P(A)$ is the proportion of times it will occur in the long-run. This is the **relative frequency** interpretation of probability.

Example 2.3. Plastic bottles for liquid laundry detergent are formed by blow molding, a manufacturing process used to create hollow plastic parts (often bottles and containers) by inflating a heated plastic tube inside a mold. Previous data from statistical process control monitoring suggests 11 percent of the bottles do not conform to specifications.

Imagine observing bottles one by one and define

$$A = \{\text{bottle does not conform to specifications}\}.$$

For each bottle observed, we note if A occurs or not. Figure 2.1 graphs what the relative frequency of the event A might look like over time when observing 1000 bottles.

2.1 Sample spaces and events

Remark: We use the phrase “random experiment” to mean an experiment which has many possible outcomes but we cannot predict with certainty which outcome will occur.

- Even if the experiment is performed the same way every time, we could still get different outcomes. “Cannot predict with certainty” means that each outcome (or collection of outcomes) has a probability associated with it.

Definition: The set of all possible outcomes for a random experiment is called the **sample space**, denoted by S . An **event** A is a subset of the sample space.

Result: If each outcome in S is equally likely (and $n_S < \infty$), then

$$P(A) = \frac{n_A}{n_S},$$

where

$$\begin{aligned} n_A &= \text{number of outcomes in } A \\ n_S &= \text{number of outcomes in } S. \end{aligned}$$

Example 2.4. Pick 3 is a three-digit number game from the South Carolina Education Lottery. We can think of playing this lottery as a random experiment with sample space

$$S = \{000, 001, 002, \dots, 998, 999\}.$$

The number of outcomes in S is $n_S = 1000$. Suppose every week I purchase three tickets with the following numbers:

$$A = \{364, 446, 540\}.$$

The probability I win is

$$P(A) = \frac{n_A}{n_S} = \frac{3}{1000} = 0.003.$$

Q: Each ticket costs \$1. Why is the payout for a winning ticket only \$500? If the game was fair, shouldn't it be closer to \$1,000?

Example 2.5. Suppose we continue to observe plastic bottles in Example 2.3 until we find the first bottle that does not conform to specifications (i.e., is “defective”). The sample space is

$$S = \{d, cd, ccd, cccd, ccccd, ccccd, \dots\},$$

where “c” and “d” represent bottles that “conform” and are “defective,” respectively.

Q: What is the probability we see the first defective bottle on the first or second bottle observed? This corresponds to the event

$$A = \{d, cd\}.$$

A: There are two problems. First of all, the sample space S does not contain a finite number of outcomes (i.e., it is countably infinite). Therefore, $n_S = +\infty$ and even writing

$$P(A) = \frac{n_A}{n_S}$$

would make no sense. Furthermore, the outcomes in S are not equally likely. It is reasonable to assign

$$P(\{d\}) = 0.11$$

for the first outcome in A because 11 percent of the bottles are thought to be defective. However, what is $P(\{cd\})$? We can't calculate this without making additional assumptions.

Definitions: The **union** of two events A and B is the event containing outcomes in A or B . The **intersection** of two events A and B is the event containing outcomes in A and B . We write

$$\begin{aligned} A \cup B &\longleftarrow \text{union ("or")} \\ A \cap B &\longleftarrow \text{intersection ("and")}. \end{aligned}$$

Example 2.6. A medical professional observes adult male patients entering an emergency room. She classifies each patient according to his blood type (AB^+ , AB^- , A^+ , A^- , B^+ , B^- , O^+ , and O^-) and whether his systolic blood pressure (SBP) is low (L), normal (N), or high (H). Imagine observing the next male patient as a random experiment. The sample space is

$$\begin{aligned} S = \{ & (AB^+, L), (AB^-, L), (A^+, L), (A^-, L), (B^+, L), (B^-, L), (O^+, L), (O^-, L), \\ & (AB^+, N), (AB^-, N), (A^+, N), (A^-, N), (B^+, N), (B^-, N), (O^+, N), (O^-, N), \\ & (AB^+, H), (AB^-, H), (A^+, H), (A^-, H), (B^+, H), (B^-, H), (O^+, H), (O^-, H) \}. \end{aligned}$$

Remarks:

- There are 8 different blood types. There are 3 different categorizations of SBP. There are $8 \times 3 = 24$ possible outcomes in the sample space, which is formed by combining the two factors. This illustrates the **multiplication rule** of counting.
- Are these 24 outcomes equally likely? Probably not. O^+ is by far the most common blood type among American males (about 38 percent). On the other hand, AB^- is rare (only about 1 percent). Similarly, most American males have either normal or high SBP; fewer have low SBP.
- Even though we have listed all possible outcomes in S , we have not specified probabilities associated with the outcomes. We cannot assign probability to events like

$$\begin{aligned} A &= \{\text{blood type with a } + \text{ rhesus status}\} \\ B &= \{\text{high SBP}\} \end{aligned}$$

without having this information.

List the outcomes in $A \cup B$ and $A \cap B$.

$$\begin{aligned} A \cup B &= \{\text{outcomes with a } + \text{ rhesus status or high SBP}\} \\ &= \{(AB^+, L), (A^+, L), (B^+, L), (O^+, L), (AB^+, N), (A^+, N), (B^+, N), (O^+, N), \\ &\quad (AB^+, H), (AB^-, H), (A^+, H), (A^-, H), (B^+, H), (B^-, H), (O^+, H), (O^-, H)\} \end{aligned}$$

$$\begin{aligned} A \cap B &= \{\text{outcomes with a } + \text{ rhesus status and high SBP}\} \\ &= \{(AB^+, H), (A^+, H), (B^+, H), (O^+, H)\} \end{aligned}$$

Discussion: The notions of union and intersection can be extended to more than two events. For example,

$$A \cup B \cup C$$

means either A , B , or C occurs. The event

$$A \cap B \cap C \cap D$$

means all four events A , B , C , and D occur. The event

$$(A \cap B) \cup (C \cap D \cap E)$$

means either $A \cap B$ or $C \cap D \cap E$ occurs. For any finite number of events A_1, A_2, \dots, A_n , we write

$$\begin{aligned} \bigcup_{i=1}^n A_i &= A_1 \cup A_2 \cup \dots \cup A_n \quad \longleftarrow \quad \text{“at least one } A_i \text{ occurs”} \\ \bigcap_{i=1}^n A_i &= A_1 \cap A_2 \cap \dots \cap A_n \quad \longleftarrow \quad \text{“each } A_i \text{ occurs”}. \end{aligned}$$

Definition: If the events A and B contain no common outcomes, we say the events are **mutually exclusive**. In this case,

$$P(A \cap B) = P(\emptyset) = 0.$$

It is not possible for A and B to both occur.

2.2 Counting techniques

Importance: When each outcome in S is equally likely, we learned

$$P(A) = \frac{n_A}{n_S},$$

where n_S and n_A count the (finite) number of outcomes in S and A , respectively. In some random experiments, counting techniques can help to determine n_S and n_A .

Multiplication rule (for counting): Suppose a random experiment involves k factors where

$$\begin{aligned} n_1 &= \text{number of outcomes for factor 1} \\ n_2 &= \text{number of outcomes for factor 2} \\ &\vdots \\ n_k &= \text{number of outcomes for factor } k. \end{aligned}$$

The total number of outcomes is

$$\prod_{i=1}^k n_i = n_1 \times n_2 \times \cdots \times n_k.$$

Example 2.7. A random experiment consists of selecting a standard South Carolina license plate which consists of 3 letters and 3 numbers. We can think of one outcome in the sample space S as having the following structure:

$$(\text{ — — — — — }).$$

Q: How many standard plates are possible; i.e., how many outcomes are in S ?

A: There are

$$n_S = 26 \times 26 \times 26 \times 10 \times 10 \times 10 = 17,576,000$$

possible outcomes.

Q: Assume each outcome in S is equally likely (e.g., license plate letters/numbers are determined at random). What is the probability a randomly selected plate contains no repeat letters and no repeat numbers?

A: Define the event

$$A = \{\text{no repeat letters/numbers}\}.$$

The number of outcomes in A is

$$n_A = 26 \times 25 \times 24 \times 10 \times 9 \times 8 = 11,232,000.$$

Therefore,

$$P(A) = \frac{n_A}{n_S} = \frac{11232000}{17576000} \approx 0.639.$$

Permutations: Suppose a random experiment involves arranging distinct objects (e.g., people, parts, locations, etc.) in order.

- With n distinct objects, there are

$$n! = n(n-1)(n-2) \times \cdots \times 2 \times 1$$

ways to permute these objects (i.e., to arrange them in a particular order).

- With n distinct objects, there are

$$P_r^n = \frac{n!}{(n-r)!}$$

ways to select r objects and then permute these.

Example 2.8. A personnel director for a corporation has hired 12 new engineers.

Q: How many ways could the engineers be assigned to 12 different offices?

A: There are

$$12! = 12 \times 11 \times 10 \times \cdots \times 2 \times 1 = 479,001,600$$

ways an assignment could be made.

Q: Suppose the director needs to select 3 engineers to fill distinct positions: team leader, consultant, and support staff member. How many ways could this be done?

A: There are

$$P_3^{12} = \frac{12!}{(12-3)!} = 12 \times 11 \times 10 = 1320$$

ways this selection could be made.

Q: In the last part, suppose there are 6 engineers from USC and 6 from Clemson. What is the probability a USC graduate is selected as the team leader and the remaining 2 positions are filled by Clemson graduates?

A: Define the event

$$A = \{\text{USC team leader and Clemson graduates for other 2 positions}\}.$$

There are a total of

$$n_S = 1320$$

ways to select 3 engineers for the distinct positions (ignoring school; this was the last part). Assuming each of these outcomes is equally likely, then all we have to do is calculate the number of outcomes in A and use

$$P(A) = \frac{n_A}{n_S}.$$

We find n_A by using the multiplication rule:

$$\begin{aligned} n_1 &= \text{number of ways to select 1 USC graduate} = 6 \\ n_2 &= \text{number of ways to select 2 Clemson graduates} = P_2^6 \end{aligned}$$

The number of outcomes in A is

$$n_A = 6 \times P_2^6 = 6 \times 30 = 180.$$

Therefore,

$$P(A) = \frac{n_A}{n_S} = \frac{180}{1320} \approx 0.136.$$

Combinations: Combinations are like permutations except the ordering of the objects doesn't matter. With n distinct objects (e.g., people, parts, locations, etc.), there are

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

ways to select r objects.

Q: In Example 2.8, how many ways are there to select 3 engineers from 12?

A: Because the order of selection doesn't matter (i.e., there are no designations of team leader or anything like that), there are

$$\binom{12}{3} = \frac{12!}{3! 9!} = \frac{12 \times 11 \times 10}{6} = 220$$

ways to select 3 engineers from 12.

Example 2.9. A bin of 50 manufactured parts contains 3 defective parts and 47 nondefective parts. A sample of 6 parts is selected at random and without replacement. That is, each part can be selected only once, and the sample is a subset of the 50 parts.

Q: What is the probability the sample contains exactly 2 defective parts?

A: Imagine the selection of 6 parts from the bin as a random experiment with outcomes in S having the form

$$(\text{---} \text{---} \text{---} \text{---} \text{---} \text{---}).$$

One possible outcome is

$$(\text{ND ND ND } \underline{\text{D}} \text{ ND ND})$$

meaning the fourth part selected was defective and the others were not.

How many outcomes are in the sample space? This is a combination question because we are simply selecting 6 parts from 50 and order doesn't matter. There are

$$n_S = \binom{50}{6} = 15,890,700$$

outcomes in S . How many outcomes are in

$$A = \{\text{sample contains 2 D and 4 ND parts}\}?$$

Use the multiplication rule with

$$\begin{aligned} n_1 &= \text{number of ways to select 2 D from 3} = \binom{3}{2} \\ n_2 &= \text{number of ways to select 4 ND from 47} = \binom{47}{4} \end{aligned}$$

The number of outcomes in A is

$$n_A = \binom{3}{2} \binom{47}{4} = 3 \times 178365 = 535,095.$$

Therefore, assuming each outcome in S is equally likely,

$$P(A) = \frac{n_A}{n_S} = \frac{535095}{15890700} \approx 0.034.$$

2.3 Axioms of probability and additive rules

Remark: Going forward, we need a formal set of rules (or axioms) which will govern how we determine probabilities in general. The following axioms are due to Kolmogorov.

Axioms: For any sample space S , assigning a probability P must satisfy

- (1) $0 \leq P(A) \leq 1$, for any event A
- (2) $P(S) = 1$
- (3) If A_1, A_2, \dots, A_n are pairwise **mutually exclusive** events, then

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i).$$

- The term “pairwise mutually exclusive” means that $A_i \cap A_j = \emptyset$, for all $i \neq j$.
- Recall the event

$$\bigcup_{i=1}^n A_i = A_1 \cup A_2 \cup \dots \cup A_n$$

means “at least one A_i occurs.”

Discussion: The first axiom guarantees that probabilities must be between 0 and 1. Events with probability 0 can never occur. Events with probability 1 must occur. Both extremes are rare in real life. The third axiom provides the mathematical basis for intuitive calculations like in the following example.

Example 2.10. In the game of craps, two fair dice are rolled initially to start the game. Here is a probability model for the sum of the two faces.

Outcome	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Q: What is the probability of rolling a “7” or an “11?”

A: The third axiom assures we can add the probabilities, that is,

$$P(\text{rolling a 7 or 11}) = P(\text{rolling a 7}) + P(\text{rolling an 11}) = \frac{6}{36} + \frac{2}{36} = \frac{8}{36} \approx 0.222.$$

This is true because

$$\{\text{rolling a 7}\} \quad \text{and} \quad \{\text{rolling an 11}\}$$

are mutually exclusive events.

Complement Rule: The complement of the event A consists of all outcomes not in A . The complement is denoted by A' . The first two axioms can be used to show

$$P(A') = 1 - P(A).$$

That is, the probability A does not occur is one minus the probability it does.

Additive Rule: If A and B are two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Of course, if A and B are mutually exclusive, then $P(A \cap B) = 0$ and

$$P(A \cup B) = P(A) + P(B).$$

This agrees with Axiom 3.

DeMorgan's Laws: If A and B are two events, then

$$\begin{aligned}(A \cup B)' &= A' \cap B' \\ (A \cap B)' &= A' \cup B'.\end{aligned}$$

Example 2.11. The probability that train 1 is on time is 0.95. The probability that train 2 is on time is 0.93. The probability that both are on time is 0.90. Define the events

$$\begin{aligned}A &= \{\text{train 1 is on time}\} \\ B &= \{\text{train 2 is on time}\}.\end{aligned}$$

We are given $P(A) = 0.95$, $P(B) = 0.93$, and $P(A \cap B) = 0.90$.

Q: What is the probability train 1 is not on time?

$$\begin{aligned}P(A') &= 1 - P(A) \\ &= 1 - 0.95 = 0.05.\end{aligned}$$

Q: What is the probability at least one train is on time?

$$\begin{aligned}P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.95 + 0.93 - 0.90 = 0.98.\end{aligned}$$

Q: What is the probability neither train is on time?

A: Note that by DeMorgan's Law,

$$\{\text{neither train on time}\} = A' \cap B' = (A \cup B)'.$$

Therefore, by the complement rule,

$$\begin{aligned}P(A \cup B)' &= 1 - P(A \cup B) \\ &= 1 - 0.98 = 0.02.\end{aligned}$$

Remark: The additive rule also applies for more than two events. For example,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

and

$$P(A \cup B \cup C \cup D) = P(A) + P(B) + P(C) + P(D) \\ - P(A \cap B) - P(A \cap C) - P(A \cap D) - P(B \cap C) - P(B \cap D) - P(C \cap D) \\ + P(A \cap B \cap C) + P(A \cap B \cap D) + P(A \cap C \cap D) + P(B \cap C \cap D) \\ - P(A \cap B \cap C \cap D)$$

for three and four events. The pattern continues, but the formula becomes increasingly impractical to work with (unless the events are mutually independent).

2.4 Conditional probability and independence

Remark: In many situations, we might want to calculate $P(A)$. However, this probability might be influenced by another event B . That is, the occurrence of B influences how we assign probability to A . This motivates conditional probability.

Definition: Let A and B be events in a sample space S with $P(B) > 0$. The **conditional probability** of A , given that B has occurred, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Similarly,

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

provided $P(A) > 0$.

Example 2.12. In a manufacturing process, 10% of the parts contain surface flaws and 25% of the parts with surface flaws are (functionally) defective parts. However, only 5% of parts without surface flaws are defective parts.

There are two events of interest here:

$$D = \{\text{part is defective}\} \\ F = \{\text{part has surface flaws}\}.$$

We are given $P(F) = 0.10$, $P(D|F) = 0.25$, and $P(D|F') = 0.05$. Therefore, how we assign probability to D depends on whether F occurs; i.e., it depends on whether the part has surface flaws (F) or not (F').

Exercise: What is $P(D)$? That is, what percentage of parts are defective overall regardless of surface flaw status?

Example 2.13. Brazilian scientists have identified a new strain of the H1N1 virus. The genetic sequence of the new strain consists of alterations in the hemagglutinin protein, making it significantly different than the usual H1N1 strain. Public health officials wish to study the population of residents in Rio de Janeiro. Suppose that in this population

- the probability of catching the usual strain is 0.10
- the probability of catching the new strain is 0.05
- the probability of catching both strains is 0.01.

Define the events

$$\begin{aligned} A &= \{\text{resident catches usual strain}\} \\ B &= \{\text{resident catches new strain}\}. \end{aligned}$$

From the information above, we have $P(A) = 0.10$, $P(B) = 0.05$, and $P(A \cap B) = 0.01$.

Q: Find the probability of catching the usual strain, given that the new strain is caught.

A: Using the definition of conditional probability,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.01}{0.05} = 0.20.$$

This means among all residents who have caught the new strain, 20% of them will also catch the usual strain. Notice how $P(A)$ and $P(A|B)$ are different. That is, the occurrence of B has changed how we assign probability to A .

Q: Find the probability of catching the new strain, given that at least one strain is caught.

A: If “at least one strain is caught,” this means $A \cup B$ has occurred. Therefore,

$$\begin{aligned} P(B|A \cup B) &= \frac{P(B \cap (A \cup B))}{P(A \cup B)} = \frac{P(B)}{P(A) + P(B) - P(A \cap B)} \\ &= \frac{0.05}{0.10 + 0.05 - 0.01} \approx 0.357. \end{aligned}$$

This means among all residents who have caught at least one strain, 35.7% of them have caught the new strain.

Multiplication Rule: If we take the conditional probability definitions

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(A \cap B)}{P(A)},$$

we see that

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ &= P(B|A)P(A). \end{aligned}$$

This is called the multiplication rule (although I’m not sure why). It gives us a way to calculate $P(A \cap B)$ when conditional probabilities are available.

- In general, we cannot determine $P(A \cap B)$ from $P(A)$ and $P(B)$ alone. We need more information or we have to make additional assumptions about A and B .

Curiosity: $P(A)$ and $P(A|B)$ are both probabilities for the event A . In the first, we look at A by itself. In the second, we incorporate knowledge that the event B has occurred. What does it mean when $P(A) = P(A|B)$? It means the knowledge gained from learning B occurred does not influence how we assign probability to the event A . This is the casual definition of **independence**.

Definition: When the occurrence or non-occurrence of B has no effect on whether or not A occurs, and vice-versa, we say the events A and B are **independent**. Mathematically, we define A and B to be independent if and only if

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

Note that if A and B are independent, then

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

and

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B)P(A)}{P(A)} = P(B).$$

Example 2.14. An electrical circuit consists of two components (e.g., resistors, capacitors, batteries, etc.). The probability the second component functions satisfactorily during its design life is 0.90. The probability at least one of the components does so is 0.96. The probability both components do so is 0.75.

Q: Do the two components function independently?

A: Define the events

$$\begin{aligned} A &= \{\text{component 1 functions}\} \\ B &= \{\text{component 2 functions}\}. \end{aligned}$$

We have $P(B) = 0.90$, $P(A \cup B) = 0.96$, and $P(A \cap B) = 0.75$. The additive rule gives

$$0.96 = P(A) + 0.90 - 0.75 \implies P(A) = 0.81.$$

However,

$$0.75 = P(A \cap B) \neq P(A)P(B) = 0.81(0.90) = 0.729.$$

Therefore, the events A and B are not independent. The two components do not function independently.

Example 2.15. In an engineering system, two components are placed in a **series**. This means the system is functional as long as both components are. Each component is functional with probability 0.95. Define the events

$$\begin{aligned} A_1 &= \{\text{component 1 is functional}\} \\ A_2 &= \{\text{component 2 is functional}\} \end{aligned}$$

so that $P(A_1) = P(A_2) = 0.95$. Because we need both components to be functional, the **system reliability** (i.e., the probability the system is functional) is $P(A_1 \cap A_2)$.

- If the components operate independently, then the system reliability is

$$P(A_1 \cap A_2) = P(A_1)P(A_2) = 0.95(0.95) = 0.9025.$$

- If the components do not operate independently; e.g., failure of one component affects the other, we can not determine the system reliability $P(A_1 \cap A_2)$ without additional knowledge or assumptions.

Remark: The notion of independence extends to more than two events. **Mutual independence** means the probability of all events in any sub-collection of A_1, A_2, \dots, A_n occurring equals the product of the probabilities of the events in the sub-collection. For example, if A_1, A_2 , and A_3 are mutually independent, then

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)P(A_2) \\ P(A_1 \cap A_3) &= P(A_1)P(A_3) \\ P(A_2 \cap A_3) &= P(A_2)P(A_3) \end{aligned}$$

considering all event sub-collections of size two. It must also be true that

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3).$$

Remark: Many random experiments can be envisioned as consisting of a sequence of n “trials” that are viewed as independent (e.g., flipping a coin 10 times). If A_i denotes the event associated with the i th trial, and the trials are mutually independent, then

$$P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i).$$

Example 2.16. Samples of 30 parts from a metal punching process are selected every hour. Each part is tested. If a part does not conform to specifications, it is sent to another location where it is repaired (or “reworked”).

Define the events

$$A_i = \{i\text{th part requires rework}\}, \quad i = 1, 2, \dots, 30.$$

Assume the 30 events A_1, A_2, \dots, A_{30} are mutually independent and $P(A_i) = p$.

Q: Under these assumptions, what is the probability at least one part requires rework?

A: Define the event

$$A = \{\text{at least one part requires rework}\} = \bigcup_{i=1}^{30} A_i.$$

The complement of A is

$$A' = \{\text{no part requires rework}\} = \bigcap_{i=1}^{30} A'_i$$

by DeMorgan's Law. Because A_1, A_2, \dots, A_{30} are mutually independent, the complements $A'_1, A'_2, \dots, A'_{30}$ are also mutually independent. Therefore,

$$P(A') = P\left(\bigcap_{i=1}^{30} A'_i\right) = \prod_{i=1}^{30} P(A'_i) = (1 - p)^{30}.$$

Finally,

$$P(A) = 1 - P(A') = 1 - (1 - p)^{30}.$$

For example, if $p = 0.01$, that is, 1% of all parts require rework, then

$$1 - (0.99)^{30} \approx 0.260.$$

Therefore, rework will be required for at least one part 26% of the time when samples of size 30 are tested.

2.5 Law of Total Probability and Bayes' Rule

Law of Total Probability: Suppose A and B are events in a sample space S . We can express A as the union of two mutually exclusive events

$$A = (A \cap B) \cup (A \cap B').$$

Therefore,

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B') \\ &= P(A|B)P(B) + P(A|B')P(B'). \end{aligned}$$

Remark: The Law of Total Probability gives us a way to calculate $P(A)$ by relying instead on the conditional probabilities $P(A|B)$ and $P(A|B')$ and the probability of a related event B . Specifically, $P(A)$ is a linear combination of the conditional probabilities $P(A|B)$ and $P(A|B')$. Note that the “weights” in the linear combination, $P(B)$ and $P(B')$, add to 1.

Example 2.17. An insurance company classifies drivers as “accident-prone” and “non-accident-prone.” The probability an accident-prone driver has an accident is 0.4. The probability a non-accident-prone driver has an accident is 0.2. The population is 30 percent accident-prone. Define the events

$$\begin{aligned} A &= \{\text{policy holder has an accident}\} \\ B &= \{\text{policy holder is accident-prone}\}. \end{aligned}$$

We are given $P(A|B) = 0.4$, $P(A|B') = 0.2$, and $P(B) = 0.3$.

Q: What is the probability a policy holder has an accident?

A: We want $P(A)$, which can be found by using the LOTP:

$$P(A) = P(A|B)P(B) + P(A|B')P(B') = 0.4(0.3) + 0.2(0.7) = 0.26.$$

Therefore, 26% of the company's policy holders will have an accident.

Q: If a policy holder has an accident, what is the probability s/he was "accident-prone?"

A: We want $P(B|A)$, which can be calculated as

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} = \frac{0.4(0.3)}{0.26} \approx 0.462.$$

Therefore, among all policy holders who had an accident, 46.2% of them are accident-prone.

Discussion: In the last part, note that if we write

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B')P(B')},$$

we obtain **Bayes' Rule** for two events.

Example 2.18. *Diagnostic testing.* A lab test is 95% effective at detecting a disease when it is present. It is 99% effective at declaring a subject negative when the subject is truly negative for the disease. Suppose 8% of the population has the disease.

Define the events

$$\begin{aligned} D &= \{\text{disease is present}\} \\ A &= \{\text{test is positive}\}. \end{aligned}$$

We are given the following information:

$$\begin{aligned} P(A|D) &= 0.95 \quad (\text{"sensitivity"}) \\ P(A'|D') &= 0.99 \quad (\text{"specificity"}) \\ P(D) &= 0.08 \quad (\text{"prevalence"}). \end{aligned}$$

Q: What is the probability a randomly selected subject will test positively?

A: We want $P(A)$. By the LOTP,

$$P(A) = P(A|D)P(D) + P(A|D')P(D') = 0.95(0.08) + 0.01(0.92) \approx 0.085.$$

Therefore, about 8.5% of the population will produce a positive test result.

Q: What is the probability a subject has the disease if his test is positive?

A: We want $P(D|A)$. By Bayes' Rule,

$$P(D|A) = \frac{P(A|D)P(D)}{P(A|D)P(D) + P(A|D')P(D')} = \frac{0.95(0.08)}{0.95(0.08) + 0.01(0.92)} \approx 0.892.$$

Therefore, among all subjects testing positively, about 89.2% of the subjects have the disease.

Remark: In general, Bayes’ Rule allows us to update probabilities on the basis of observed information. In Example 2.18, the “observed information” we get to see is the test result.

Prior probability	Test result		Posterior probability
$P(D) = 0.08$	\longrightarrow	A	\longrightarrow $P(D A) \approx 0.892$
$P(D) = 0.08$	\longrightarrow	A'	\longrightarrow $P(D A') \approx 0.004$

2.6 Introduction to random variables

Terminology: A **random variable** is a variable whose value is determined by chance.

- By convention, we denote random variables by upper case letters towards the end of the alphabet; e.g., W , X , Y , Z , etc. The authors of the textbook for this course (Montgomery and Runger, 2018) favor the use of X .

Motivation: In Example 2.16, we considered samples of 30 parts from a metal punching process. Conceptualizing this as random experiment, the sample space can be written as

$$S = \{(0, 0, 0, \dots, 0), (1, 0, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (1, 1, 1, \dots, 1)\},$$

where “1” denotes a part that requires rework and “0” denotes a part that does not. There are

$$n_S = 2^{30} = 1,073,741,824$$

outcomes in this sample space! Examples like this illustrate why working with random variables is easier. After all, the line technician or quality control engineer is probably only interested in how many of the parts will require rework. If we let

$$X = \text{number of parts requiring rework (out of 30)},$$

then we no longer have to work with an unwieldy sample space with over a billion outcomes. We can instead work with events of the form

$$\{X = x\},$$

and designate probabilities for $P(X = x)$, for $x = 0, 1, 2, \dots, 30$. This is much easier.

Terminology: Random variables generally break down into one of two types.

- If a random variable X can have only a finite (or countable) number of values, we say it is **discrete**.
 - For example, the random variable

$$X = \text{number of parts requiring rework (out of 30)}$$

is discrete because there are 31 possible values $x = 0, 1, 2, \dots, 30$.

- If it makes more sense that X has values in an interval of numbers, we say X is **continuous**.
 - Measurements like part diameters (cm), temperature (deg C), energy expended (kcal), time (days, years, etc.), and distance (miles) are all best regarded as continuous random variables.

Terminology: The set of all values a random variable X can have is called its **support**.

Example 2.19. Classify the following random variables as discrete or continuous and specify the support of each:

- V = number of unbroken eggs in a randomly selected carton (dozen)
- W = pH of an aqueous solution
- X = length of time between accidents at a factory
- Y = whether or not you pass this class
- Z = number of aircraft arriving tomorrow at CAE.

- The random variable V is **discrete**. It can have values in

$$\{0, 1, 2, \dots, 12\}.$$

- The random variable W is **continuous**. It can have values in

$$\mathbb{R} = \{-\infty < w < \infty\}.$$

With most solutions, it is more likely that W is not negative (although this is possible) and not larger than, say, 15 (a very reasonable upper bound).

- The random variable X is **continuous**. It can have values in

$$\mathbb{R}^+ = \{x > 0\}.$$

The key point here is that a time cannot be negative. In theory, it is possible that X can be very large.

- The random variable Y is **discrete**. It can have values in

$$\{0, 1\},$$

where I have arbitrarily labeled “1” for passing and “0” for failing. Random variables that can assume exactly two values (e.g., 0, 1) are **binary**.

- The random variable Z is **discrete**. It can have values in

$$\mathbb{N} = \{0, 1, 2, 3, \dots\}.$$

I have allowed for the possibility of a very large number of aircraft arriving.

3 Discrete Distributions

3.1 Probability mass functions

Recall: A random variable X is **discrete** if it can have a finite or countable number of values.

Terminology: The **probability mass function** (pmf) of a discrete random variable X tells us two things:

1. the values X can have
2. a probability $p_X(x) = P(X = x)$ for each value of x .

The pmf of X ,

$$p_X(x) = P(X = x),$$

describes the **distribution** of X . It tells us which values of x are possible (the support of X) and how to assign probabilities to these values.

Example 3.1. An automobile paint factory has 5 filling lines (lines where cans are filled with paint) which are in continuous operation. During a 24-hour time period, mechanics record

$X =$ the number of lines which require maintenance.

Here is the distribution of X , described by its pmf:

x	0	1	2	3	4	5
$p_X(x)$	0.60	0.10	0.16	0.05	0.06	0.03

The support of X is

$$\{0, 1, 2, 3, 4, 5\}.$$

Also,

- the probabilities $p_X(x) = P(X = x)$ are all between 0 and 1
- the probabilities add to 1.

These two things must be true for any pmf (otherwise, we say the pmf is not valid).

Q: What is the probability no lines require maintenance during a 24-hour time period?

$$P(X = 0) = 0.60.$$

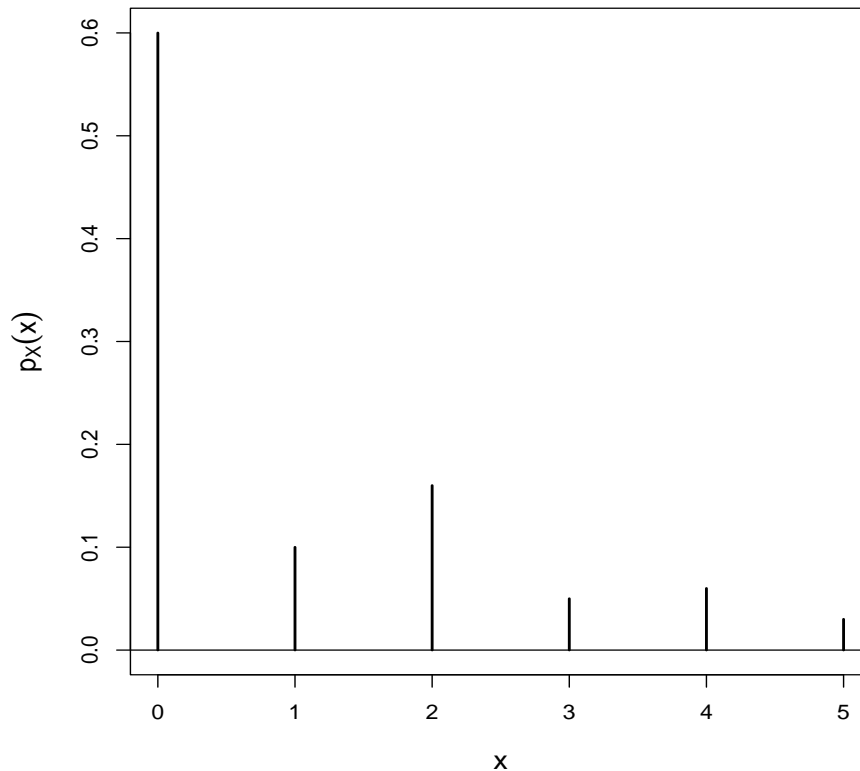


Figure 3.1: Probability mass function of X in Example 3.1.

Q: Upper management is contacted whenever 3 or more lines require maintenance during a 24-hour time period. What percentage of days will this occur?

$$\begin{aligned} P(X \geq 3) &= P(X = 3) + P(X = 4) + P(X = 5) \\ &= 0.05 + 0.06 + 0.03 = 0.14. \end{aligned}$$

Upper management will be contacted on 14% of the days.

Terminology: The **cumulative distribution function** (cdf) of a discrete random variable X gives probabilities of the form

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

for any real number x .

- When X is a discrete random variable, the cdf $F_X(x)$ is a step function.
- The range of $F_X(x)$ is

$$0 \leq F_X(x) \leq 1.$$

This makes sense because $F_X(x)$ is a probability.

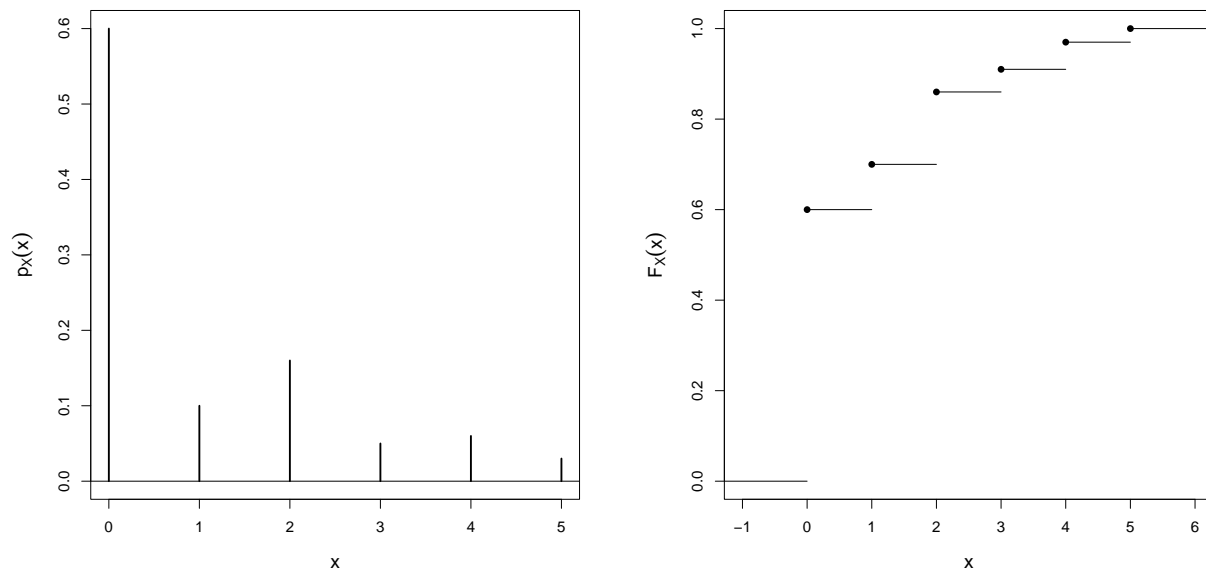


Figure 3.2: Left: Probability mass function (pmf) of X in Example 3.1. Right: Cumulative distribution function (cdf) of X .

Note: A discrete random variable's cdf cumulates (adds up) probability as we move from left to right on the pmf. The cdf of X in Example 3.1, shown above, can be written as

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 0.60, & 0 \leq x < 1 \\ 0.70, & 1 \leq x < 2 \\ 0.86, & 2 \leq x < 3 \\ 0.91, & 3 \leq x < 4 \\ 0.97, & 4 \leq x < 5 \\ 1, & x \geq 5. \end{cases}$$

Example 3.1 (continued). Answer the following questions by using both the pmf and cdf in Figure 3.2.

Q: What is the probability there are at most 2 lines that require maintenance?

A: We want $P(X \leq 2)$.

PMF:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.60 + 0.10 + 0.16 = 0.86$$

CDF:

$$P(X \leq 2) = F_X(2) = 0.86.$$

Q: What is the probability there are at least 4 lines that require maintenance?

A: We want $P(X \geq 4)$.

PMF:

$$P(X \geq 4) = P(X = 4) + P(X = 5) = 0.06 + 0.03 = 0.09$$

CDF:

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - F_X(3) = 1 - 0.91 = 0.09.$$

Q: What is the probability there is exactly 1 line that requires maintenance?

A: We want $P(X = 1)$.

PMF:

$$P(X = 1) = 0.10$$

CDF:

$$P(X = 1) = P(X \leq 1) - P(X \leq 0) = F_X(1) - F_X(0) = 0.70 - 0.60 = 0.10.$$

3.2 Mean and variance

Terminology: Suppose X is a discrete random variable with pmf $p_X(x)$. The **expected value** of X is

$$\mu = E(X) = \sum_{\text{all } x} xp_X(x).$$

The expected value of a discrete random variable X is a weighted average of the possible values of X . Each value x is weighted by its probability $p_X(x)$. In statistical applications, $\mu = E(X)$ is called the **mean** or **population mean**.

Example 3.1 (continued). We examined the distribution of X , the number of filling lines which require maintenance during a 24-hour period. The probability mass function (pmf) of X is

x	0	1	2	3	4	5
$p_X(x)$	0.60	0.10	0.16	0.05	0.06	0.03

The expected value of X is

$$\begin{aligned} E(X) &= \sum_{\text{all } x} xp_X(x) \\ &= 0(0.60) + 1(0.10) + 2(0.16) + 3(0.05) + 4(0.06) + 5(0.03) = 0.96. \end{aligned}$$

Interpretations:

- $E(X)$ is the “center of gravity” of a discrete random variable’s pmf. It’s the location on the horizontal axis where $p_X(x)$ would balance if it were made of solid material.

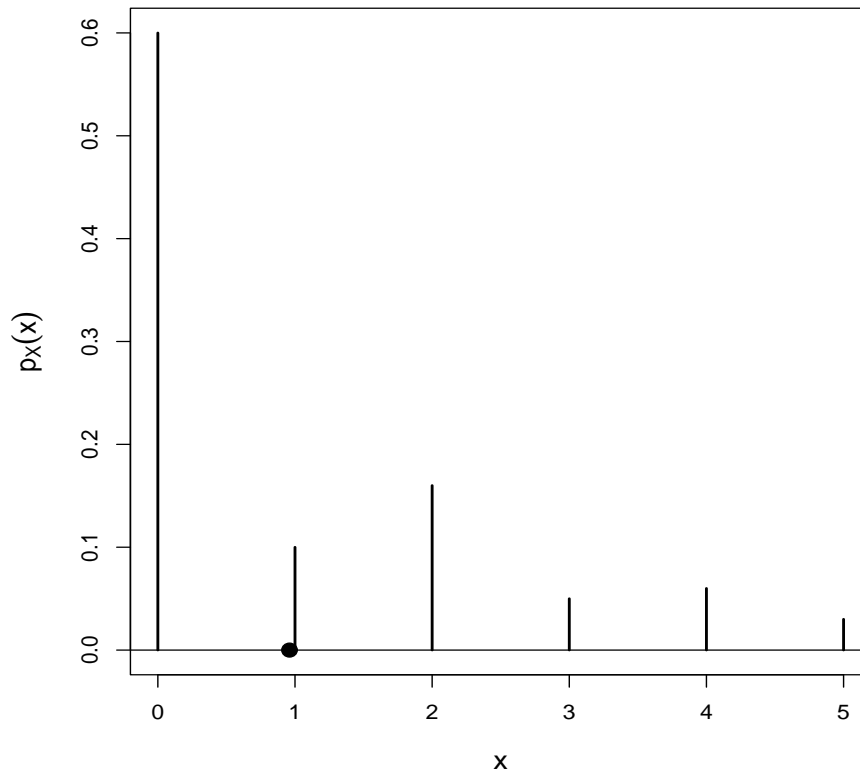


Figure 3.3: Probability mass function of X in Example 3.1. The expected value $E(X) = 0.96$ is shown by a solid circle.

- $E(X)$ is a “long-run average.” That is, if we were to observe the value of X for many 24-hour periods in Example 3.1, the average of these observations should be “close” to $E(X)$. The more observations there are, the closer this average will be to $E(X)$. For example, the R code below simulates 365 measurements of X in Example 3.1 and averages them.

```
> options(digits=3) # control number of significant digits presented
> n = 365
> x = c(0,1,2,3,4,5)
> prob = c(0.60,0.10,0.16,0.05,0.06,0.03)
> lines = sample(x,n,replace=TRUE,prob=prob)
> mean(lines)
[1] 0.934
```

Result: Suppose X is a discrete random variable with pmf $p_X(x)$ and g is any function. Then $g(X)$ is also a random variable and its expectation (mean) is

$$E[g(X)] = \sum_{\text{all } x} g(x)p_X(x).$$

Linearity rules:

- (a) $E(c) = c$, for any constant c
- (b) $E[cg(X)] = cE[g(X)]$, for any constant c
- (c) The expectation of the sum is the sum of the expectations; i.e.,

$$E[g_1(X) + g_2(X) + \cdots + g_k(X)] = E[g_1(X)] + E[g_2(X)] + \cdots + E[g_k(X)].$$

We say the expectation $E(\cdot)$ is a **linear operator**, that is, it preserves the operations of addition and scalar multiplication.

Example 3.2. In a one-hour period, the number of gallons of toxic waste produced at a local plant, say X , has the following pmf:

x	0	1	2	3
$p_X(x)$	0.2	0.3	0.3	0.2

This pmf is shown in Figure 3.4 (next page).

Q: What is $E(X)$?

$$E(X) = \sum_{\text{all } x} xp_X(x) = 0(0.2) + 1(0.3) + 2(0.3) + 3(0.2) = 1.5.$$

Therefore, we would expect 1.5 gallons of toxic waste to be produced per hour.

Q: Disposing toxic waste requires careful handling and adherence to regulations to protect human health and the environment. The cost C (in \$100s) to dispose X gallons of waste per hour is a quadratic function of X , specifically,

$$C = 3 - 1.4X + 4.6X^2.$$

What is the expected cost of disposal $E(C)$ in a one-hour period?

A: We can use the linearity rules stated above. Let's find $E(X^2)$ first:

$$E(X^2) = \sum_{\text{all } x} x^2 p_X(x) = 0^2(0.2) + 1^2(0.3) + 2^2(0.3) + 3^2(0.2) = 3.3.$$

Therefore,

$$\begin{aligned} E(C) = E(3 - 1.4X + 4.6X^2) &= 3 - 1.4E(X) + 4.6E(X^2) \\ &= 3 - 1.4(1.5) + 4.6(3.3) = 16.08. \end{aligned}$$

The expected hourly cost of toxic waste disposal is \$1,608.00.

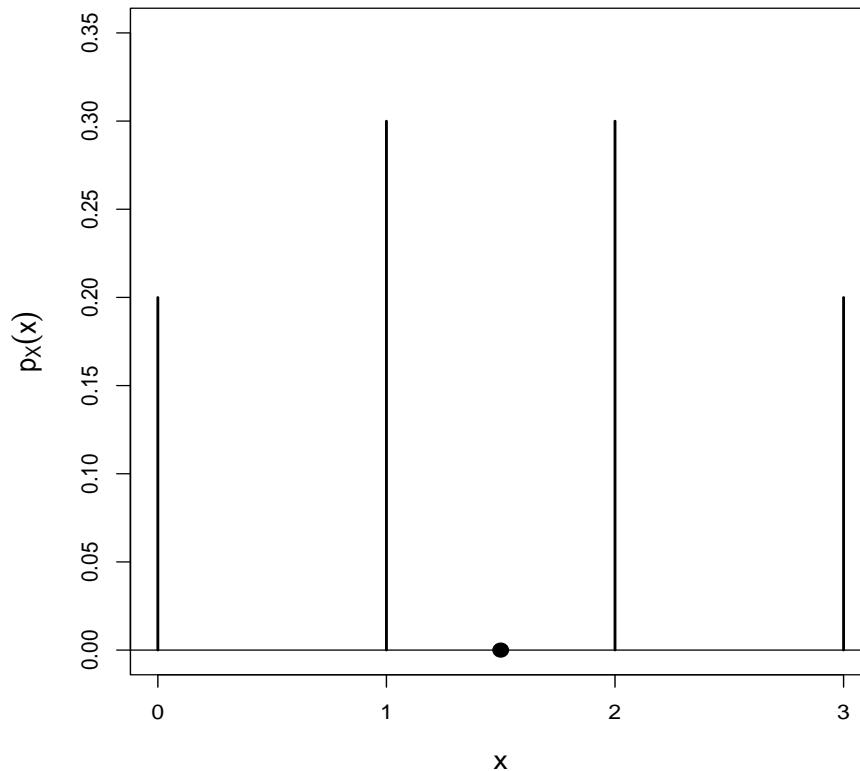


Figure 3.4: Probability mass function of X in Example 3.2. The expected value $E(X) = 1.5$ is shown by a solid circle.

Terminology: Suppose X is a discrete random variable with pmf $p_X(x)$ and mean $\mu = E(X)$. The **variance** of X is

$$\begin{aligned}\sigma^2 = V(X) &= E[(X - \mu)^2] \\ &= \sum_{\text{all } x} (x - \mu)^2 p_X(x).\end{aligned}$$

The **standard deviation** of X is the positive square root of the variance:

$$\sigma = \sqrt{\sigma^2} = \sqrt{V(X)}.$$

Example 3.2 (continued). Recall the pmf for the number of gallons of toxic waste produced (per hour) is

x	0	1	2	3
$p_X(x)$	0.2	0.3	0.3	0.2

Q: Find $V(X)$ and the standard deviation of X .

Using the definition of the variance, we have

$$\begin{aligned} V(X) &= \sum_{\text{all } x} (x - \mu)^2 p_X(x) \\ &= (0 - 1.5)^2(0.2) + (1 - 1.5)^2(0.3) + (2 - 1.5)^2(0.3) + (3 - 1.5)^2(0.2) = 1.05. \end{aligned}$$

The standard deviation of X is

$$\sigma = \sqrt{V(X)} = \sqrt{1.05} \approx 1.025.$$

Interpreting the variance and standard deviation:

1. Whereas $E(X)$ measures the “center” of a distribution, the variance $V(X)$ measures the “spread,” that is, how spread out the values of X are about the mean.
2. The larger $V(X)$ is, the more spread (variability) in the distribution of X .
3. The variance $V(X) \geq 0$. The only time $V(X) = 0$ is when X has a **degenerate distribution**; i.e., all the probability mass is at one point.
4. $V(X)$ is measured in (units)² and σ is measured in original units.
5. From the definition,

$$V(X) = E[(X - \mu)^2]$$

is the expected squared distance between X and the mean. The standard deviation σ is (roughly) the expected distance between X and the mean.

Linear functions: Suppose X is a discrete random variable with pmf $p_X(x)$ and mean $\mu = E(X)$. Suppose a and b are constants. The mean and variance of the linear function $aX + b$ are

$$\begin{aligned} E(aX + b) &= aE(X) + b \\ V(aX + b) &= a^2V(X). \end{aligned}$$

The expectation result follows from the linearity properties associated with $E(\cdot)$. The variance result is new.

- Taking $a = 0$, we see that $V(b) = 0$ for any constant b . This makes sense—the variance is a measure of variability for a random variable; a constant does not vary. This also means that additive shifts of b (to the left or right) do not affect the spread in the distribution of X .
- Taking $b = 0$, we see that $V(aX) = a^2V(X)$. Multiplicative constants increase the variance when $|a| > 1$ and decrease the variance when $0 < |a| < 1$.

Example 3.3. Patient responses to a generic drug to control pain are scored on 5-point scale (1 = lowest pain level; 5 = highest pain level). In a certain population of patients, the pmf of the response X is given by

x	1	2	3	4	5
$p_X(x)$	0.38	0.27	0.18	0.11	0.06

This pmf is shown in Figure 3.5 (next page, left).

Q: Find the expected value and variance of X .

A: The expected value is

$$\begin{aligned} E(X) &= \sum_{\text{all } x} xp_X(x) \\ &= 1(0.38) + 2(0.27) + 3(0.18) + 4(0.11) + 5(0.06) = 2.2. \end{aligned}$$

In this application, it would be appropriate to call $E(X) = \mu = 2.2$ the **population mean**. It is the mean pain response for all patients in the population under study.

The variance is

$$\begin{aligned} V(X) &= \sum_{\text{all } x} (x - \mu)^2 p_X(x) \\ &= (1 - 2.2)^2(0.38) + (2 - 2.2)^2(0.27) + (3 - 2.2)^2(0.18) \\ &\quad + (4 - 2.2)^2(0.11) + (5 - 2.2)^2(0.06) = 1.5. \end{aligned}$$

Similarly, it would be appropriate to call $V(X) = \sigma^2 = 1.5$ the **population variance**. It is the variance associated with the pain responses for all patients in the population.

Q: Find the mean and variance of $Y = 2X - 1$.

A: The mean of $Y = 2X - 1$ is

$$E(2X - 1) = 2E(X) - 1 = 2(2.2) - 1 = 3.4.$$

The variance of $Y = 2X - 1$ is

$$V(2X - 1) = 4V(X) = 4(1.5) = 6.$$

These are the mean and variance associated with the pmf of $Y = 2X - 1$:

y	1	3	5	7	9
$p_Y(y)$	0.38	0.27	0.18	0.11	0.06

This pmf is shown in Figure 3.5 (next page, right).

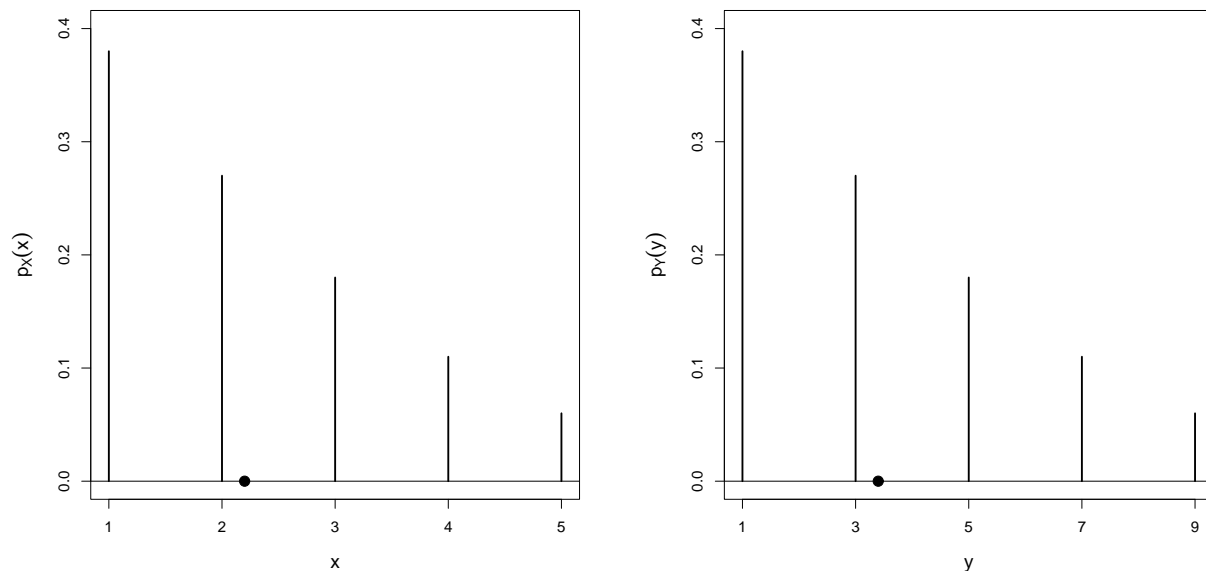


Figure 3.5: Left: Probability mass function of X in Example 3.3. Right: Probability mass function of $Y = 2X - 1$. The means $E(X) = 2.2$ and $E(Y) = 3.4$ are shown by solid circles.

Variance computing formula: Suppose X is a random variable (discrete or continuous) with mean $\mu = E(X)$. An alternative way to find $V(X)$ is by using

$$V(X) = E(X^2) - [E(X)]^2.$$

This formula is easy to remember and can make calculations easier. It also reminds us that

$$E(X^2) \neq [E(X)]^2.$$

Some students are tempted to write $E(X^2) = [E(X)]^2$, but this is not true! Provided X does not have a degenerate distribution (where all the probability is at one value), note that

$$V(X) > 0 \implies E(X^2) > [E(X)]^2.$$

Example 3.3 (continued). For the pmf

x	1	2	3	4	5
$p_X(x)$	0.38	0.27	0.18	0.11	0.06

we have

$$\begin{aligned} E(X^2) &= \sum_{\text{all } x} x^2 p_X(x) \\ &= 1^2(0.38) + 2^2(0.27) + 3^2(0.18) + 4^2(0.11) + 5^2(0.06) = 6.34. \end{aligned}$$

Using the variance computing formula,

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 \\ &= 6.34 - (2.2)^2 = 1.5. \end{aligned}$$

This is the same answer we got for $V(X)$ when we used the definition of variance.

3.3 Binomial distribution

Bernoulli trials: Many random experiments can be envisioned as consisting of a sequence of “trials,” where

1. each trial results in a “success” or a “failure” (only 2 outcomes are possible)
2. the trials are independent (the result on one trial is not affected by the results from other trials)
3. the probability of success p is the same on every trial.

Examples:

- When circuit boards used in the manufacture of laptops are tested, one percent of the boards are found to be defective.
 - circuit board = “trial”
 - defective board is observed = “success”
 - $p = 0.01$
- Ninety-eight percent of all air traffic radar signals are correctly interpreted the first time they are transmitted.
 - radar signal = “trial”
 - signal is correctly interpreted = “success”
 - $p = 0.98$
- Albino rats used to study the hormonal regulation of a metabolic pathway are injected with a drug that inhibits body synthesis of protein. Twenty percent of all rats will die before the study is complete.
 - rat = “trial”
 - dies before study is over = “success”
 - $p = 0.2$

- During her WNBA tenure, Caitlyn Clark’s free throw percentage is 88.7%.
 - free throw = “trial”
 - made free throw = “success”
 - $p = 0.887$

Definition: A **binomial distribution** arises when we observe a fixed number of Bernoulli trials:

$$\begin{aligned} n &= \text{number of trials} \\ X &= \text{number of successes (out of } n\text{)}. \end{aligned}$$

If the Bernoulli trial assumptions hold, then the probability mass function (pmf) of X is given by the formula

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \dots, n \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim b(n, p)$, where p is the probability of success on any one trial.

Example 3.4. In an agricultural study, 40 percent of all plots respond to a certain treatment. In this context, we interpret

- plot = “trial”
- plot responds to treatment = “success”
- $p = 0.4$

Four plots are observed. If the Bernoulli trial assumptions hold (independent plots, same response probability for each plot), then

$$X = \text{number of plots responding to treatment} \sim b(n = 4, p = 0.4).$$

This pmf is shown in Figure 3.6 (next page). Here are all the pmf calculations:

$$\begin{aligned} P(X = 0) &= \binom{4}{0} (0.4)^0 (0.6)^4 = 0.1296 \\ P(X = 1) &= \binom{4}{1} (0.4)^1 (0.6)^3 = 0.3456 \\ P(X = 2) &= \binom{4}{2} (0.4)^2 (0.6)^2 = 0.3456 \\ P(X = 3) &= \binom{4}{3} (0.4)^3 (0.6)^1 = 0.1536 \\ P(X = 4) &= \binom{4}{4} (0.4)^4 (0.6)^0 = 0.0256. \end{aligned}$$

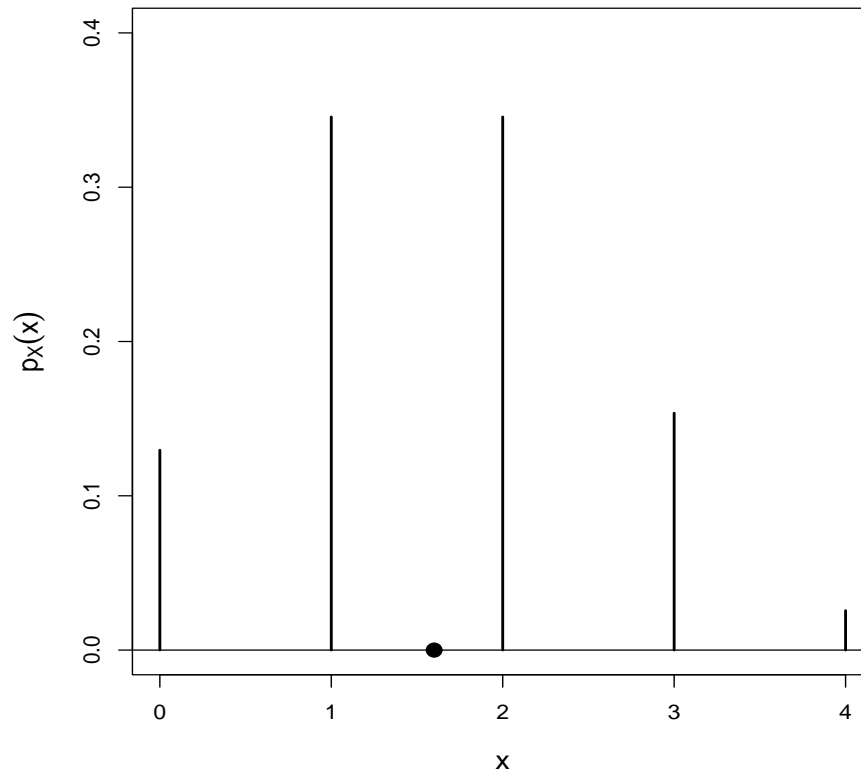


Figure 3.6: Probability mass function of $X \sim b(4, 0.4)$ in Example 3.4. The expected value $E(X) = 1.6$ is shown by a solid circle.

Putting these in tabular form (like before), we have

x	0	1	2	3	4
$p_X(x)$	0.1296	0.3456	0.3456	0.1536	0.0256

MEAN/VARIANCE: If $X \sim b(n, p)$, then

$$\begin{aligned} E(X) &= np \\ V(X) &= np(1 - p). \end{aligned}$$

Example 3.4 (continued). The expected number of plots which respond to treatment is

$$E(X) = 4(0.4) = 1.6 \text{ plots.}$$

The variance is

$$V(X) = 4(0.4)(0.6) = 0.96 \text{ (plots)}^2.$$

The standard deviation is

$$\sigma = \sqrt{0.96} \approx 0.98 \text{ plots.}$$

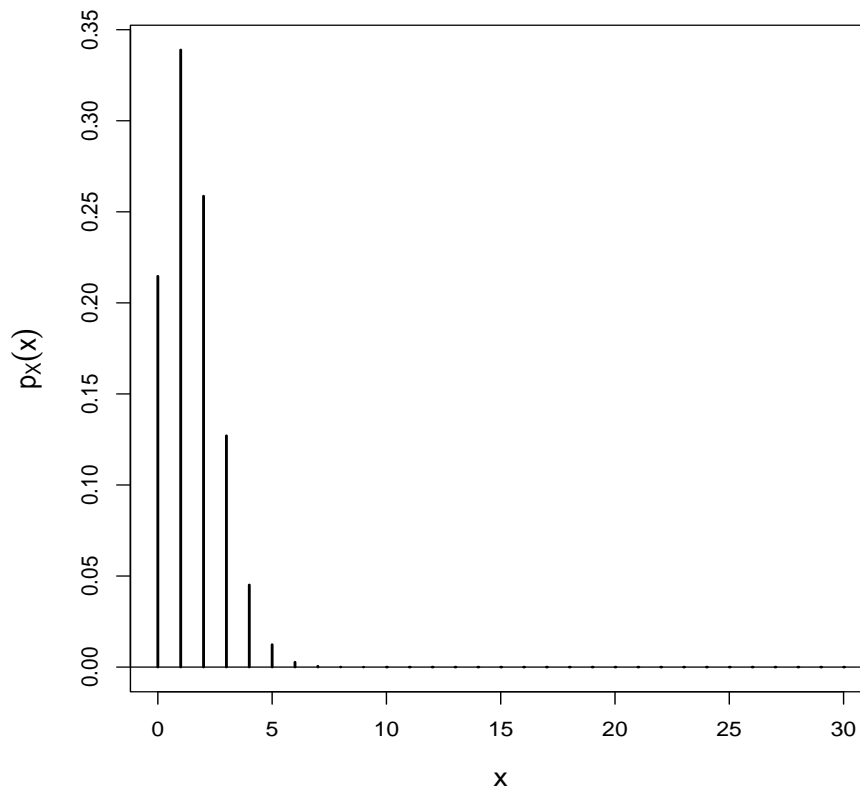


Figure 3.7: Probability mass function of $X \sim b(30, 0.05)$ in Example 3.5.

Example 3.5. A computer’s power supply unit (PSU) is hardware that converts high-voltage alternating current (from a wall outlet) into lower-voltage direct current power which is needed by a computer’s components (e.g., CPU, etc.). A manufacturer claims “no more than 5 percent” of its power supply units need servicing during their warranty period. We can interpret

- PSU = “trial”
- PSU needs service during warranty period = “success”
- $p = 0.05$ (in fact, the manufacturer claims p is no larger than 0.05).

Technicians access a sample of 30 units and simulate their usage during the warranty period. If the Bernoulli trial assumptions hold (independent units, same probability of needing service for each unit), then

$$X = \text{number of PSUs needing service during warranty period} \sim b(n = 30, p = 0.05).$$

This pmf is shown in Figure 3.7 (above).

Q: Among the 30 units tested, what is the probability the technicians find 5 or more PSUs requiring service during the warranty period?

A: We want $P(X \geq 5)$. We could calculate

$$P(X = 5) + P(X = 6) + P(X = 7) + \cdots + P(X = 29) + P(X = 30),$$

but this would require using the binomial pmf formula 26 times and adding up the results. It is easier to write

$$\begin{aligned} P(X \geq 5) &= 1 - P(X \leq 4) \\ &= 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3) - P(X = 4) \end{aligned}$$

which requires using the binomial pmf formula only 5 times. Note that

$$\begin{aligned} P(X = 0) &= \binom{30}{0} (0.05)^0 (0.95)^{30} \approx 0.2146 \\ P(X = 1) &= \binom{30}{1} (0.05)^1 (0.95)^{29} \approx 0.3389 \\ P(X = 2) &= \binom{30}{2} (0.05)^2 (0.95)^{28} \approx 0.2586 \\ P(X = 3) &= \binom{30}{3} (0.05)^3 (0.95)^{27} \approx 0.1270 \\ P(X = 4) &= \binom{30}{4} (0.05)^4 (0.95)^{26} \approx 0.0451. \end{aligned}$$

Therefore,

$$P(X \geq 5) \approx 1 - 0.2146 - 0.3389 - 0.2586 - 0.1270 - 0.0451 = 0.0158.$$

Under the $b(30, 0.05)$ model, it is unlikely the technicians would find 5 or more PSUs requiring service during the warranty period. This would occur only in 1.58% of the samples of size 30 tested.

Discussion: What if the technicians *did* observe 5 PSUs which required service during the warranty period? What might be true?

BINOMIAL R CODE: Suppose $X \sim b(n, p)$.

$p_X(x) = P(X = x)$	$F_X(x) = P(X \leq x)$
<code>dbinom(x,n,p)</code>	<code>pbinom(x,n,p)</code>

```
> options(digits=4)
> dbinom(2,30,0.05) # P(X=2)
[1] 0.2586
> 1-pbinom(4,30,0.05) # 1-P(X<=4)
[1] 0.01564
```


Note: Another way in R to calculate $P(X \geq 5)$ in this example would be to use the code

```
> sum(dbinom(5:30,30,0.05))
[1] 0.01564
```

The `dbinom(5:30,30,0.05)` command creates a vector containing the binomial pmf probabilities $p_X(5), p_X(6), \dots, p_X(30)$. The `sum` command adds them.

3.4 Geometric and negative binomial distributions

Note: Both the geometric and negative binomial distributions arise from observing Bernoulli trials.

Definition: A **geometric distribution** arises when we continue to observe Bernoulli trials until the first success occurs. Specifically, define

X = number of trials to observe the 1st success.

If the Bernoulli trial assumptions hold, then the probability mass function (pmf) of X is given by the formula

$$p_X(x) = \begin{cases} (1-p)^{x-1}p, & x = 1, 2, 3, \dots, \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \text{geom}(p)$, where p is the probability of success on any one trial.

Example 3.6. An EPA engineer is tasked with testing water specimens from lakes in northeast Georgia. In this region, each specimen has a 20 percent chance of containing a particular organic pollutant. We interpret

- specimen = “trial”
- specimen contains the pollutant = “success”
- $p = 0.2$

Define

X = number of specimens tested to find the first one containing the pollutant.

If the Bernoulli trial assumptions hold (independent specimens, each specimen has the same probability of containing the pollutant), then $X \sim \text{geom}(p = 0.2)$. This pmf is shown in Figure 3.8 (next page).

Q: What is the probability the engineer finds the first polluted specimen on the 5th specimen tested?

A: We want

$$P(X = 5) = (0.8)^4(0.2) \approx 0.082.$$

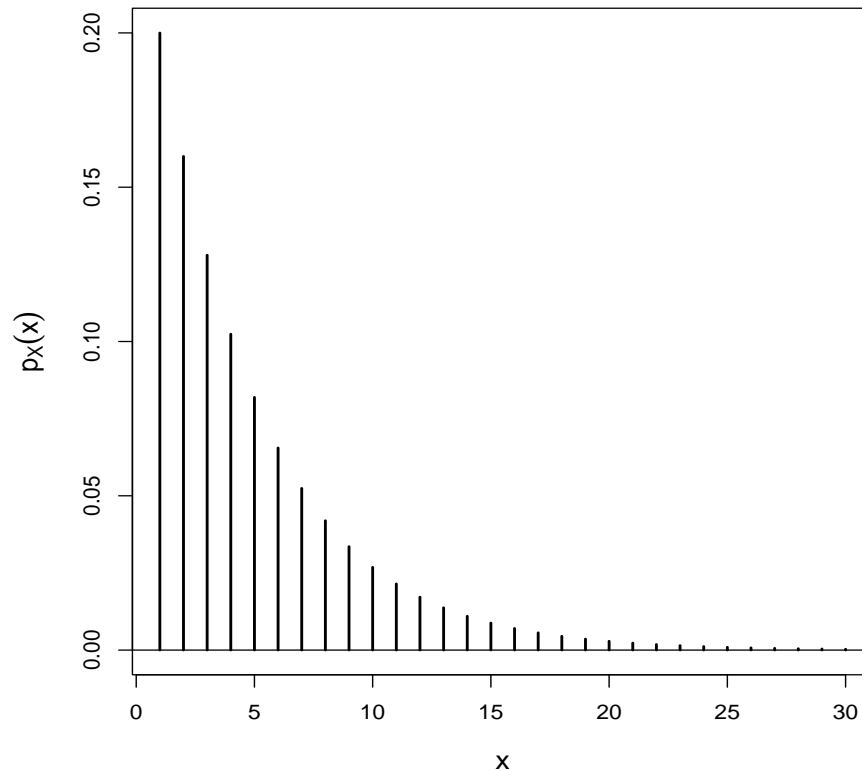


Figure 3.8: Probability mass function of $X \sim \text{geom}(p = 0.2)$ in Example 3.6.

Q: What is the probability the first polluted specimen is found before the 4th specimen is tested?

A: We want

$$\begin{aligned}
 P(X \leq 3) &= P(X = 1) + P(X = 2) + P(X = 3) \\
 &= (0.8)^0(0.2) + (0.8)^1(0.2) + (0.8)^2(0.2) = 0.488.
 \end{aligned}$$

GEOMETRIC R CODE: Suppose $X \sim \text{geom}(p)$.

$p_X(x) = P(X = x)$	$F_X(x) = P(X \leq x)$
<code>dgeom(x-1,p)</code>	<code>pgeom(x-1,p)</code>

```

> options(digits=3)
> dgeom(5-1,0.2) # P(X=5)
[1] 0.0819
> pgeom(3-1,0.2) # P(X<=3)
[1] 0.488

```

Definition: A **negative binomial distribution** arises when we continue to observe Bernoulli trials until the r th success occurs. Specifically, define

X = number of trials to observe the r th success.

If the Bernoulli trial assumptions hold, then the probability mass function (pmf) of X is given by the formula

$$p_X(x) = \begin{cases} \binom{x-1}{r-1} (1-p)^{x-r} p^r, & x = r, r+1, r+2, \dots, \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \text{nib}(r, p)$, where p is the probability of success on any one trial.

Note: Some authors call r the **waiting parameter** in the distribution, because we are “waiting” to observe the r th success. Of course, if $r = 1$, then the negative binomial pmf reduces to the geometric pmf.

Example 3.7. At an automotive paint factory, 25 percent of all batches sent to the lab for chemical analysis do not conform to specifications. This might occur for batches that were not prepared properly in the mixing and/or grinding stages. In this situation, we interpret

- batch = “trial”
- batch does not conform = “success”
- $p = 0.25$

Q: What is the probability the 2nd nonconforming batch is found on the 6th batch tested?

A: We are “waiting” until we find the 2nd nonconforming batch ($r = 2$). If the Bernoulli trial assumptions hold (independent batches, same probability of nonconforming for each batch), then

$$\begin{aligned} X &= \text{the number of batches tested to find the second nonconforming} \\ &\sim \text{nib}(r = 2, p = 0.25). \end{aligned}$$

This pmf is shown in Figure 3.9 (next page).

We want

$$P(X = 6) = \binom{5}{1} (0.75)^4 (0.25)^2 \approx 0.099.$$

Q: What is the probability we need to observe 20 or more batches to find the 2nd nonconforming batch?

A: We want

$$\begin{aligned} P(X \geq 20) &= 1 - P(X \leq 19) \\ &= 1 - P(X = 2) - P(X = 3) - \dots - P(X = 19) \\ &= 1 - \underbrace{\binom{1}{1} (0.75)^0 (0.25)^2 - \binom{2}{1} (0.75)^1 (0.25)^2 - \dots - \binom{18}{1} (0.75)^{17} (0.25)^2}_{\text{there are 18 terms here}} \approx 0.031. \end{aligned}$$

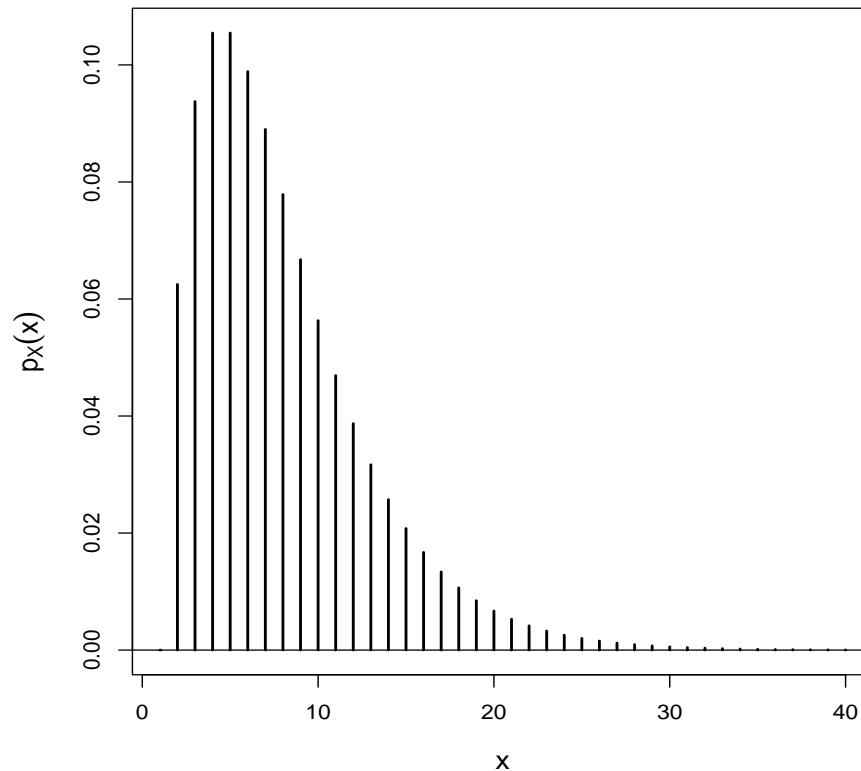


Figure 3.9: Probability mass function of $X \sim \text{nib}(2, 0.25)$ in Example 3.7.

NEGATIVE BINOMIAL R CODE: Suppose $X \sim \text{nib}(r, p)$.

$p_X(x) = P(X = x)$	$F_X(x) = P(X \leq x)$
<code>dnbinom(x-r,r,p)</code>	<code>pnbinom(x-r,r,p)</code>

```
> dnbinom(6-2,2,0.25) # P(X=6)
[1] 0.0989
> 1-pnbinom(19-2,2,0.25) # 1-P(X<=19)
[1] 0.031
```

MEAN/VARIANCE: If $X \sim \text{nib}(r, p)$, then

$$E(X) = \frac{r}{p}$$

$$V(X) = \frac{r(1-p)}{p^2}.$$

Letting $r = 1$ in the formulas above gives $E(X)$ and $V(X)$ for $X \sim \text{geom}(p)$.

3.5 Hypergeometric distribution

Setting: Consider a population of N objects and suppose each object belongs to one of two classes: Class 1 or Class 2. For example, the objects might be people (infected/not), parts (defective/not), plots of land (respond to treatment/not), etc. We have

$$\begin{aligned} N &= \text{total number of objects} \\ K &= \text{number of objects in Class 1} \\ N - K &= \text{number of objects in Class 2.} \end{aligned}$$

A sample of n objects is taken from the population at random and without replacement (after an object is selected, it is not replaced).

Definition: In the setting above, a **hypergeometric distribution** arises when we observe

$$X = \text{number of Class 1 objects in the sample (out of } n\text{)}.$$

The probability mass function (pmf) of X is given by the formula

$$p_X(x) = \begin{cases} \frac{\binom{K}{x} \binom{N-K}{n-x}}{\binom{N}{n}}, & x \leq K \text{ and } n-x \leq N-K \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \text{hyper}(N, n, K)$, where N is the population size, n is the sample size, and K is the number of Class 1 objects in the population.

Discussion: The motivation for the hypergeometric distribution should remind us of the underlying framework for the binomial; i.e., we record the number of Class 1 objects (“successes”) out of n (“trials”). The difference here is that

- the population size N is finite
- sampling is done without replacement.

To understand further, suppose

$$p = \frac{K}{N} = \text{proportion of Class 1 objects in the population.}$$

Because sampling from the population is done **without replacement**, the value of p changes from trial to trial. This violates the Bernoulli trial assumptions, so technically the binomial model does not apply. However, if the population size N is “large,” the $\text{hyper}(N, n, K)$ distribution and the $b(n, p = K/N)$ distribution should be very close to each other even when one samples without replacement. Of course, if one samples from the population **with replacement**, then $p = K/N$ remains fixed and hence the binomial model applies regardless of how large N is.

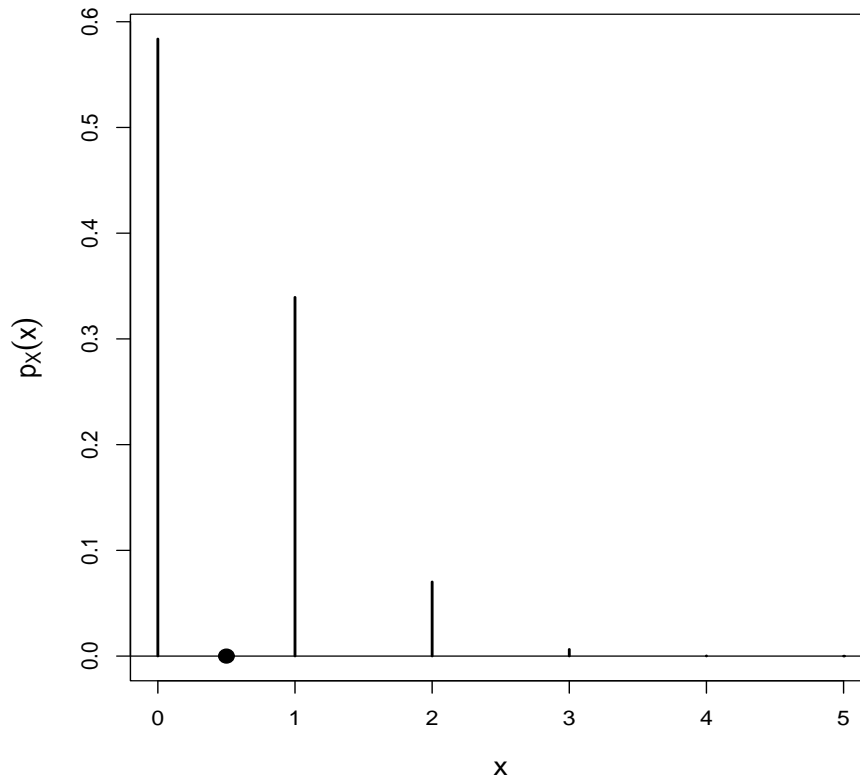


Figure 3.10: Probability mass function of $X \sim \text{hyper}(100, 5, 10)$ in Example 3.8. The expected value $E(X) = 0.5$ is shown by a solid circle.

Example 3.8. A supplier ships parts to a company in lots of 100 parts. The company has an acceptance sampling plan which adopts the following rule:

“....sample 5 parts at random and without replacement. If there are no defectives in the sample, accept the entire lot; otherwise, reject the entire lot.”

Suppose a lot contains 10 defective parts and 90 non-defective parts (this information would usually not be known to the company). Define

X = number of defective parts in the sample (out of 5)

so that $X \sim \text{hyper}(N = 100, n = 5, K = 10)$. This pmf is shown in Figure 3.10 (above).

Q: Following the rule above, what is the probability the company accepts the lot?

A: We want $P(X = 0)$. The lot will be accepted only when there are no defectives in the sample.

$$P(X = 0) = \frac{\binom{10}{0} \binom{90}{5}}{\binom{100}{5}} \approx 0.584.$$

Q: What is the probability there are at most 2 defectives in the sample?

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \frac{\binom{10}{0} \binom{90}{5}}{\binom{100}{5}} + \frac{\binom{10}{1} \binom{90}{4}}{\binom{100}{5}} + \frac{\binom{10}{2} \binom{90}{3}}{\binom{100}{5}} \approx 0.993. \end{aligned}$$

HYPERGEOMETRIC R CODE: Suppose $X \sim \text{hyper}(N, n, K)$.

$$\frac{p_X(x) = P(X = x)}{\text{dhyper}(x, K, N - K, n)} \quad \frac{F_X(x) = P(X \leq x)}{\text{phyper}(x, K, N - K, n)}$$

```
> options(digits=3)
> dhyper(0,10,100-10,5) # P(X=0)
[1] 0.584
> phyper(2,10,100-10,5) # P(X<=2)
[1] 0.993
```

MEAN/VARIANCE: If $X \sim \text{hyper}(N, n, K)$, then

$$\begin{aligned} E(X) &= n \left(\frac{K}{N} \right) \\ V(X) &= n \left(\frac{K}{N} \right) \left(1 - \frac{K}{N} \right) \left(\frac{N - n}{N - 1} \right). \end{aligned}$$

Note that these formulas are similar to $E(X) = np$ and $V(X) = np(1 - p)$ when $X \sim b(n, p)$ and $p = K/N$ is the proportion of Class 1 objects in the population. The extra term in the variance formula is called the **finite population correction factor** (it adjusts for the fact one is sampling from a finite population—not one which is regarded to be infinite in size).

Example 3.8 (continued). The expected number of defective parts sampled is

$$E(X) = 5 \left(\frac{10}{100} \right) = 0.5 \text{ parts.}$$

The variance is

$$V(X) = 5 \left(\frac{10}{100} \right) \left(1 - \frac{10}{100} \right) \left(\frac{95}{99} \right) = 0.432 \text{ (parts)}^2.$$

The standard deviation is

$$\sigma = \sqrt{0.432} \approx 0.657 \text{ parts.}$$

3.6 Poisson distribution

Relevance: The Poisson distribution is the most common probability distribution when modeling **counts** such as

- the number of customers entering a post office per hour
- the number of insurance claims received per day
- the number of machine breakdowns per month
- the number of severe weather events per year
- the number of raw material defects per square foot
- the number of airborne aerosol particles per cubic inch.

Definition: A **Poisson distribution** arises when we are counting the number of “occurrences” over a unit interval of time (e.g., hour, day, month, year etc.) or a unit region of space (e.g., square foot, cubic inch, etc.). These occurrences must obey the following postulates:

- P1. The number of occurrences in non-overlapping intervals of time (or regions of space) are independent.
- P2. The probability of an occurrence is proportional to the length of the interval of time (or area/volume of the region of space).
- P3. The probability of 2 or more occurrences in a sufficiently small interval of time (or region of space) is zero.

Define

X = number of occurrences in a unit interval of time (or region of space).

If the Poisson postulate assumptions hold, then the probability mass function (pmf) of X is given by the formula

$$p_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, \dots \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \text{Poisson}(\lambda)$, where λ is the mean number of occurrences per unit interval of time or region of space.

MEAN/VARIANCE: If $X \sim \text{Poisson}(\lambda)$, then

$$\begin{aligned} E(X) &= \lambda \\ V(X) &= \lambda. \end{aligned}$$

This is a unique feature of the Poisson distribution—its mean and variance are equal.

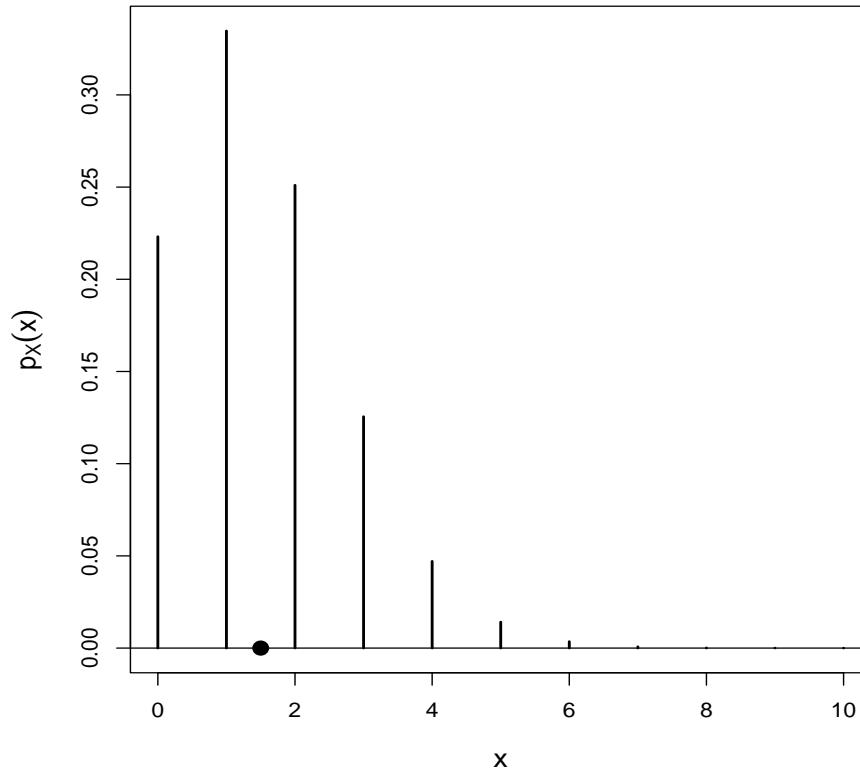


Figure 3.11: Probability mass function of $X \sim \text{Poisson}(1.5)$ in Example 3.9. The expected value $E(X) = 1.5$ is shown by a solid circle.

Example 3.9. In a certain region in the northeast US, the number of severe weather events per year X is assumed to have a Poisson distribution with mean $\lambda = 1.5$. The pmf of X is shown in Figure 3.11 (above).

Q: What is the probability there are exactly 2 severe weather events in a given year?

$$P(X = 2) = \frac{(1.5)^2 e^{-1.5}}{2!} \approx 0.251.$$

Q: What is the probability there are more than 3 severe weather events in a given year?

A: We want $P(X \geq 4)$.

$$\begin{aligned}
 P(X \geq 4) &= 1 - P(X \leq 3) \\
 &= 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3) \\
 &= 1 - \frac{(1.5)^0 e^{-1.5}}{0!} - \frac{(1.5)^1 e^{-1.5}}{1!} - \frac{(1.5)^2 e^{-1.5}}{2!} - \frac{(1.5)^3 e^{-1.5}}{3!} \approx 0.066.
 \end{aligned}$$

POISSON R CODE: Suppose $X \sim \text{Poisson}(\lambda)$.

$$\frac{p_X(x) = P(X = x)}{\text{dpois}(x, \lambda)} \quad \frac{F_X(x) = P(X \leq x)}{\text{ppois}(x, \lambda)}$$

```
> options(digits=3)
> dpois(2,1.5) # P(X=2)
[1] 0.251
> 1-ppois(3,1.5) # 1-P(X<=3)
[1] 0.0656
```

Example 3.9 (continued). A local company in this region buys an insurance policy in the event severe weather shuts down business. The policy pays nothing for the first severe weather event of the year but pays \$25,000 for each one thereafter, until the end of the year.

Q: Calculate the expected amount paid to the company under this policy during a one-year period.

A: First note that if $X = 0$ or $X = 1$, then the company receives nothing according to the policy. It is only when there are 2 or more severe weather events does a payout occur, and this payout is \$25,000 for each severe weather event. Therefore, the payout when viewed as a function of X is

$$g(X) = \begin{cases} 0, & X = 0, 1 \\ 25000(X - 1), & X = 2, 3, 4, \dots \end{cases}$$

and we want to calculate $E[g(X)]$. From the definition of expectation, we have

$$\begin{aligned} E[g(X)] &= \sum_{x=0}^{\infty} g(x) \frac{(1.5)^x e^{-1.5}}{x!} \\ &= 0 \times \frac{(1.5)^0 e^{-1.5}}{0!} + 0 \times \frac{(1.5)^1 e^{-1.5}}{1!} + \sum_{x=2}^{\infty} 25000(x-1) \frac{(1.5)^x e^{-1.5}}{x!} \\ &= 25000 \sum_{x=2}^{\infty} (x-1) \frac{(1.5)^x e^{-1.5}}{x!} \\ &= 25000 \left[\sum_{x=0}^{\infty} (x-1) \frac{(1.5)^x e^{-1.5}}{x!} - (1-1) \times \frac{(1.5)^1 e^{-1.5}}{1!} - (0-1) \times \frac{(1.5)^0 e^{-1.5}}{0!} \right]. \end{aligned}$$

Note that

$$\sum_{x=0}^{\infty} (x-1) \frac{(1.5)^x e^{-1.5}}{x!} = E(X-1) = E(X) - 1 = 1.5 - 1 = 0.5.$$

Therefore,

$$E[g(X)] = 25000(0.5 - 0 + e^{-1.5}) \approx 18078.25.$$

The expected payout to the company during a one-year period is \$18078.25.

4 Continuous Distributions

4.1 Probability density functions

Recall: A random variable X is **continuous** if, at least in theory, it can have any value in an interval of numbers. For example,

- $X = \text{pH of an aqueous solution} \rightarrow \mathbb{R} = \{-\infty < x < \infty\}$
- $X = \text{length of time (days) between accidents at a factory} \rightarrow \mathbb{R}^+ = \{x > 0\}$
- $X = \text{proportion of parts which require rework} \rightarrow [0, 1] = \{0 \leq x \leq 1\}$
- $X = \text{current (mA) measured in a thin copper wire} \rightarrow [4.9, 5.1] = \{4.9 \leq x \leq 5.1\}$
- $X = \text{diameter (mm) of a hole drilled in sheet metal} \rightarrow (12.5, \infty) = \{x > 12.5\}.$

Important: Assigning probabilities to events with continuous random variables is different than in the discrete case. We do not assign probability to specific values like $x = 2$ as we might in discrete models. Instead, we assign probabilities to events which are **intervals** like $1 < x < 3$, acknowledging that X can have any value between 1 and 3.

Terminology: The **probability density function** (pdf) of a continuous random variable X is a function $f_X(x)$ which has the following characteristics:

1. $f_X(x) \geq 0 \rightarrow f_X(x)$ is nonnegative.
2. the area under any pdf is equal to 1, that is,

$$\int_{-\infty}^{\infty} f_X(x) dx = 1.$$

Result: If X is a continuous random variable with pdf $f_X(x)$, then

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

Probabilities with continuous random variables are found by integrating the pdf.

Example 4.1. The amount of loss/damage (in millions of dollars) due to catastrophic weather is a continuous random variable X with pdf

$$f_X(x) = \begin{cases} \frac{3000}{(10+x)^4}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is shown in Figure 4.1 (next page).

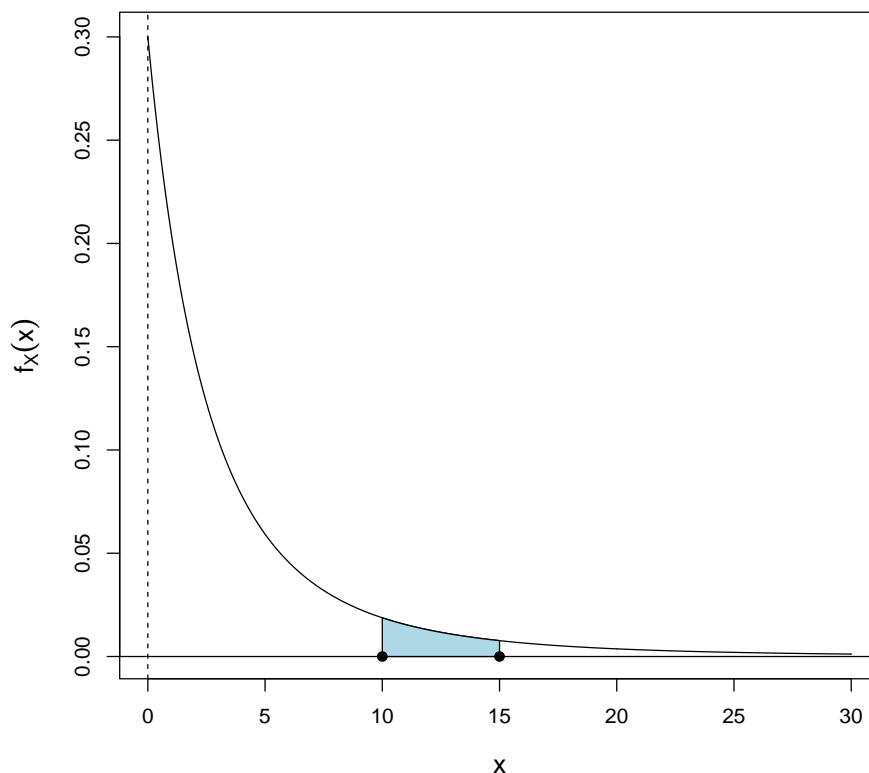


Figure 4.1: Probability density function of X in Example 4.1. The shaded area under the curve is $P(10 < X < 15)$.

Q: What is the probability the amount of loss/damage is between 10 and 15 million dollars?

A: We want

$$P(10 < X < 15) = \int_{10}^{15} f_X(x) dx = \int_{10}^{15} \frac{3000}{(10+x)^4} dx.$$

To do this integral, let

$$u = 10 + x \implies du = dx.$$

With this u -substitution (noting the change in limits), the last integral equals

$$\int_{20}^{25} \frac{3000}{u^4} du = 3000 \int_{20}^{25} \frac{1}{u^4} du = 3000 \left(-\frac{1}{3} \frac{1}{u^3} \right) \Big|_{20}^{25} = 1000 \left(\frac{1}{20^3} - \frac{1}{25^3} \right) = 0.061.$$

Therefore,

$$P(10 < X < 15) = 0.061.$$

That is, 6.1% of all loss/damage amounts from catastrophic weather will be between 10 and 15 million dollars.

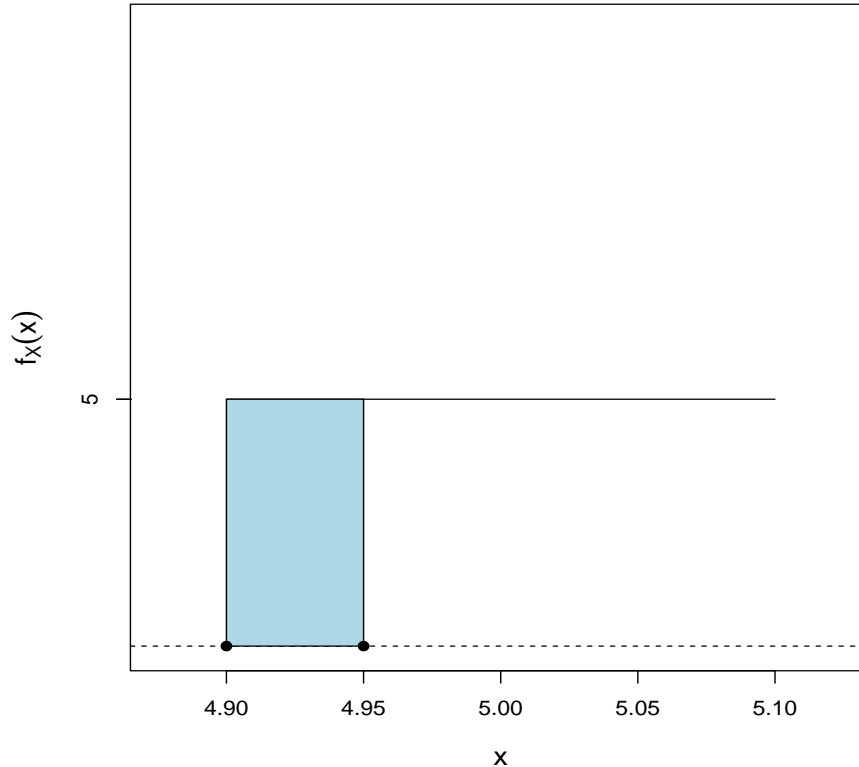


Figure 4.2: Probability density function of X in Example 4.2. The shaded area under the curve is $P(X < 4.95)$.

Integration in R: One-dimensional integrals can be found numerically using the `integrate` function in R. In Example 4.1,

```
> pdf <- function(x){3000/(10+x)^4}
> integrate(pdf,lower=10,upper=15) # P(10<X<15)
0.061 with absolute error < 6.8e-16
```

The absolute error arises from the adaptive quadrature method used to perform numerical integration (it is generally very small for “well behaved” functions we are integrating).

Example 4.2. Let the continuous random variable X denote the current measured in a thin copper wire (in milliamperes, mA). Assume the range (support) of X is $[4.9, 5.1]$ mA and the pdf of X is

$$f_X(x) = \begin{cases} 5, & 4.9 \leq x \leq 5.1 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is shown in Figure 4.2 (above).

Q: What is the probability the current measurement is less than 4.95 mA?

A: We want

$$P(X < 4.95) = \int_{4.9}^{4.95} f_X(x) dx = \int_{4.9}^{4.95} 5 dx = 5 \left(x \Big|_{4.9}^{4.95} \right) = 5(4.95 - 4.9) = 0.25.$$

That is, 25% of all current measurements will be less than 4.95 mA.

Remark: In Example 4.2, the pdf is constant across the range (support) of X . We call this a **uniform distribution**. Probability is assigned equally to intervals of the same size across the support of X .

Example 4.3. UPS ships millions of packages every month in a specific 1-ft³ packing container. Define

X = amount of space occupied in this container (in ft³).

The pdf of X is given by

$$f_X(x) = \begin{cases} 90x^8(1-x), & 0 < x < 1 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is shown in Figure 4.3 (next page).

Q: What is the probability a package will be filled at 0.9 ft³ or more?

A: We want

$$\begin{aligned} P(X \geq 0.9) &= \int_{0.9}^1 f_X(x) dx = \int_{0.9}^1 90x^8(1-x) dx \\ &= 90 \int_{0.9}^1 (x^8 - x^9) dx \\ &= 90 \left(\frac{x^9}{9} - \frac{x^{10}}{10} \right) \Big|_{0.9}^1 = 90 \left(\frac{1}{9} - \frac{1}{10} - \frac{0.9^9}{9} + \frac{0.9^{10}}{10} \right) \approx 0.264. \end{aligned}$$

That is, approximately 26.4% of all containers will be filled at 0.9 ft³ or more.

R: Here is how to calculate this probability in R:

```
> options(digits=3)
> pdf <- function(x){90*x^8*(1-x)}
> integrate(pdf,lower=0.9,upper=1) # P(X>=0.9)
0.264 with absolute error < 2.9e-15
```

Discussion: Discrete distributions (last chapter) assign positive probability to specific points. Calculating probabilities with continuous distributions is done by integration. We integrate a continuous random variable's pdf $f_X(x)$ over the range defined by the event of interest. In continuous distributions, all single points are assigned zero probability.

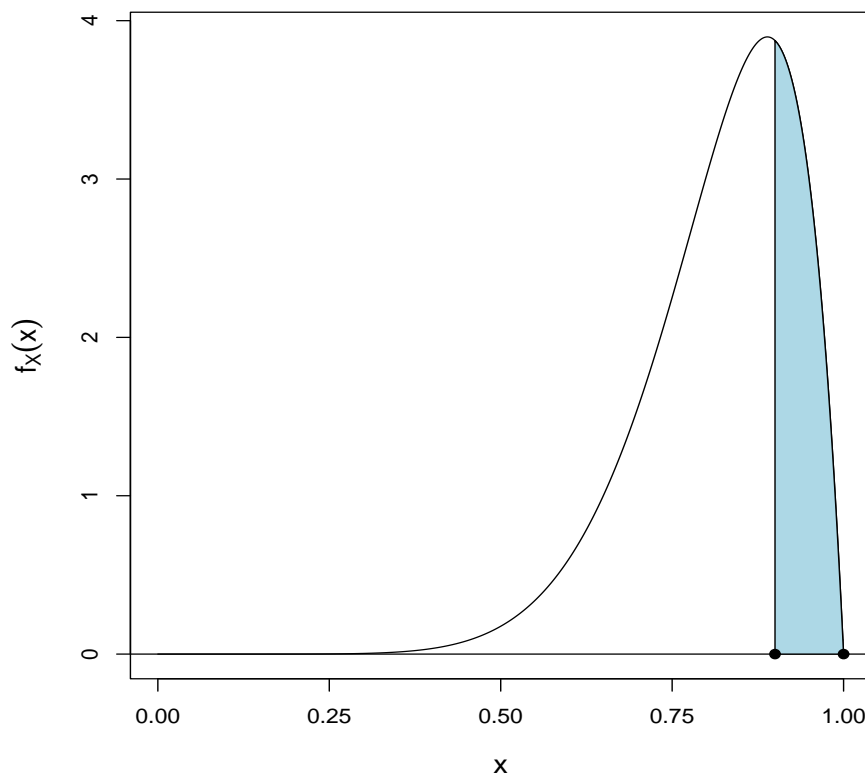


Figure 4.3: Probability density function of X in Example 4.3. The shaded area under the curve is $P(X \geq 0.9)$.

Why? This makes sense if you think about it using calculus. Suppose a is a number in the support of X . The probability

$$P(X = a) = \int_a^a f_X(x) dx = 0.$$

The area under $f_X(x)$ above a single point is always zero. This highlights the salient difference between discrete and continuous random variables. In discrete models, specific points have positive probability. In continuous models, they don't. It follows that

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

and each one equals

$$\int_a^b f_X(x) dx.$$

The endpoints simply don't matter when X is continuous. Of course, this is not true in discrete distributions.

Terminology: The **cumulative distribution function** (cdf) of a continuous random variable X gives probabilities of the form

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t)dt$$

for any real number x .

- When X is a continuous random variable, the cdf $F_X(x)$ is a continuous function.
- Applying the Fundamental Theorem of Calculus (part 1), it follows that

$$\frac{d}{dx}F_X(x) = \frac{d}{dx} \int_{-\infty}^x f_X(t)dt = f_X(x).$$

That is, differentiating a continuous random variable's cdf produces the pdf.

Example 4.1 (continued). The amount of loss/damage (in millions of dollars) due to catastrophic weather is a continuous random variable X with pdf

$$f_X(x) = \begin{cases} \frac{3000}{(10+x)^4}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

The cdf of X is

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - \left(\frac{10}{10+x}\right)^3, & x \geq 0. \end{cases}$$

The pdf and cdf are shown side by side in Figure 4.4 (next page). We previously calculated

$$P(10 < X < 15) = \int_{10}^{15} f_X(x)dx = \int_{10}^{15} \frac{3000}{(10+x)^4}dx = 0.061$$

by integrating the pdf over $(10, 15)$. We can also get this from the cdf:

$$\begin{aligned} P(10 < X < 15) &= P(X < 15) - P(X < 10) \\ &= F_X(15) - F_X(10) = \left[1 - \left(\frac{10}{10+15}\right)^3\right] - \left[1 - \left(\frac{10}{10+10}\right)^3\right] = 0.061. \end{aligned}$$

This example illustrates the following general result for **continuous** random variables:

$$P(a < X < b) = \int_a^b f_X(x)dx = F_X(b) - F_X(a),$$

which, in essence, applies the Fundamental Theorem of Calculus (part 2).

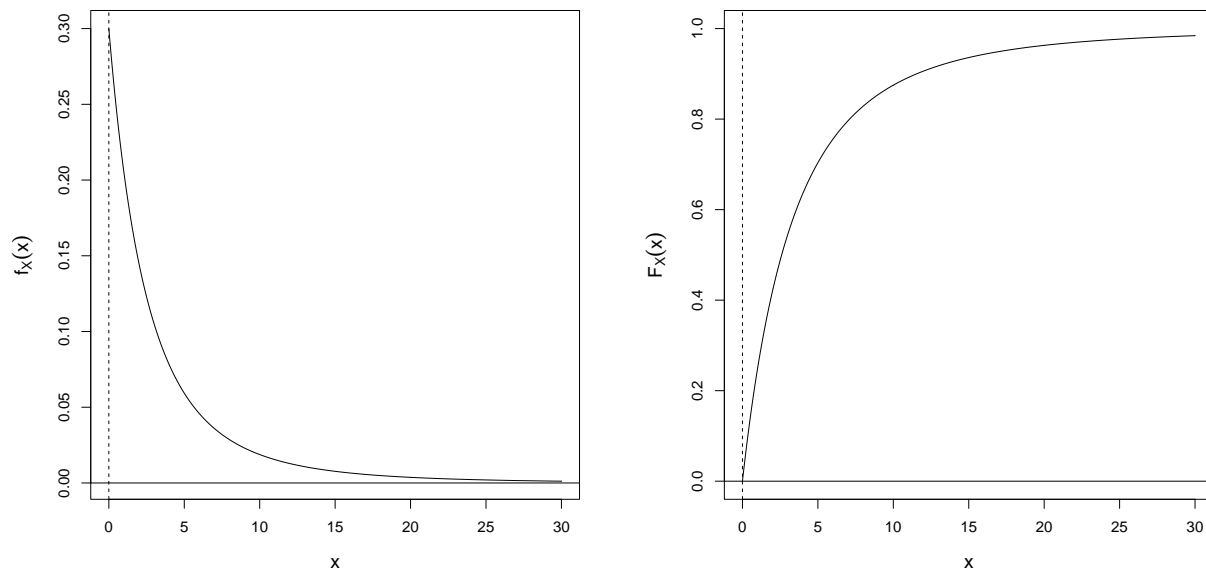


Figure 4.4: Left: Probability density function (pdf) of X in Example 4.1. Right: Cumulative distribution function (cdf) of X .

Examples 4.2 and 4.3 (continued). The pdfs and cdfs are

$$4.2. \quad f_X(x) = \begin{cases} 5, & 4.9 \leq x \leq 5.1 \\ 0, & \text{otherwise} \end{cases} \implies F_X(x) = \begin{cases} 0, & x < 4.9 \\ 5x - 24.5, & 4.9 \leq x \leq 5.1 \\ 1, & x > 5.1 \end{cases}$$

$$4.3. \quad f_X(x) = \begin{cases} 90x^8(1-x), & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases} \implies F_X(x) = \begin{cases} 0, & x \leq 0 \\ 10x^9 - 9x^{10}, & 0 < x < 1 \\ 1, & x \geq 1 \end{cases}.$$

These functions are shown in Figures 4.5 and 4.6 (next page).

- In Example 4.2, using the cdf gives

$$P(X < 4.95) = F_X(4.95) = 5(4.95) - 24.5 = 0.25.$$

- In Example 4.3, using the cdf gives

$$P(X \geq 0.9) = 1 - P(X \leq 0.9) = 1 - F_X(0.9) = 1 - [10(0.9)^9 - 9(0.9)^{10}] \approx 0.264.$$

Note: These are the same answers we got by using the pdfs (and integrating). The lesson here is that knowing a random variable's cdf can greatly simplify our work. For upcoming “named” continuous distributions, like the exponential, gamma, normal, and others, the corresponding cdfs are available in R.

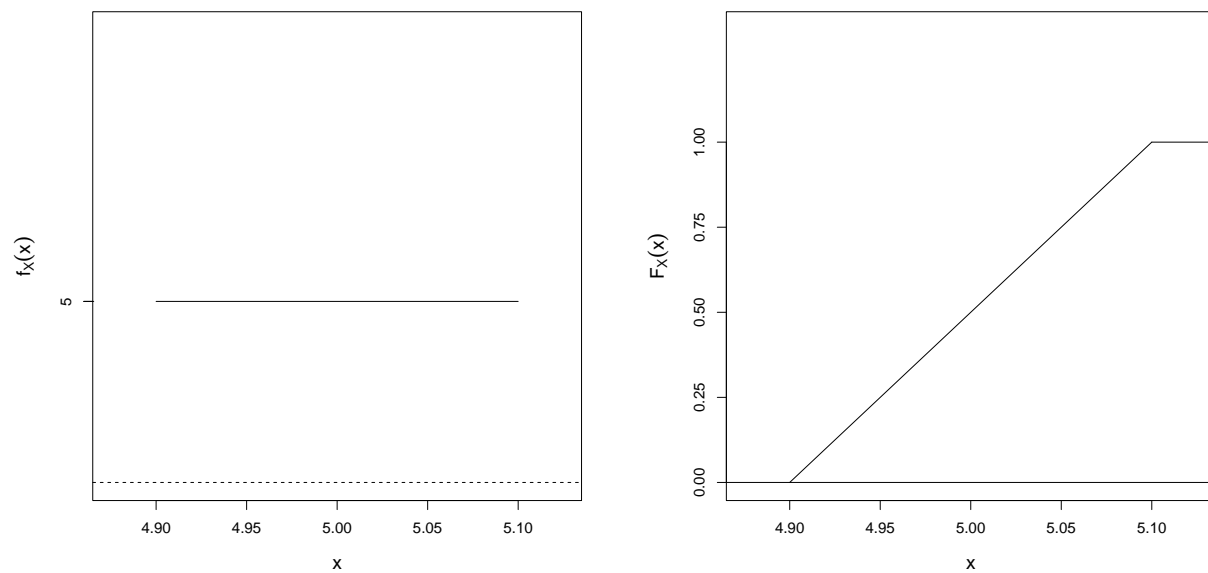


Figure 4.5: Left: Probability density function (pdf) of X in Example 4.2. Right: Cumulative distribution function (cdf) of X .

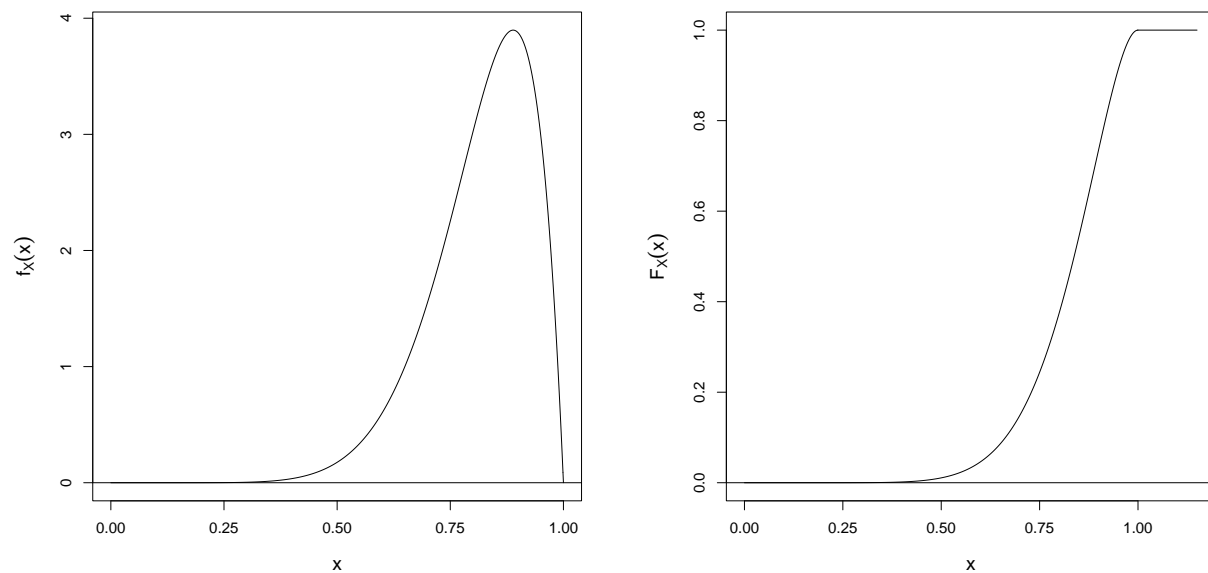


Figure 4.6: Left: Probability density function (pdf) of X in Example 4.3. Right: Cumulative distribution function (cdf) of X .

4.2 Mean, variance, and percentiles

Terminology: Suppose X is a continuous random variable with pdf $f_X(x)$. The **expected value** of X is

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

We interpret $E(X)$ the same way as we did in the discrete case:

- a center of gravity or “balance point” on the pdf
- a “long-run average.”

In statistical applications, $\mu = E(X)$ is called the **mean** or **population mean**.

Remark: The limits of the integral in the definition $E(X)$ above, while technically correct, will always be the lower and upper limits corresponding to the nonzero part of the pdf.

Example 4.4. Let the continuous random variable X denote the diameter of a hole drilled in a sheet metal component. The target diameter is 12.5 millimeters. However, random disturbances to the drilling process result in larger diameters. Historical data show the distribution of X can be modeled by the pdf

$$f_X(x) = \begin{cases} 20e^{-20(x-12.5)}, & x > 12.5 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is shown in Figure 4.7 (next page).

Q: Find the mean diameter $E(X)$.

A: We calculate

$$E(X) = \int_{12.5}^{\infty} x f_X(x) dx = \int_{12.5}^{\infty} 20x e^{-20(x-12.5)} dx = 20e^{20(12.5)} \int_{12.5}^{\infty} x e^{-20x} dx.$$

To do the last integral, use integration by parts with

$$\begin{aligned} u &= x & du &= dx \\ dv &= e^{-20x} dx & v &= -\frac{1}{20} e^{-20x} \end{aligned}$$

so that

$$\begin{aligned} \int_{12.5}^{\infty} x e^{-20x} dx &= \left(-\frac{x}{20} e^{-20x} \right) \Big|_{12.5}^{\infty} - \int_{12.5}^{\infty} -\frac{1}{20} e^{-20x} dx \\ &= \frac{12.5}{20} e^{-20(12.5)} - 0 + \frac{1}{20} \left(-\frac{1}{20} e^{-20x} \right) \Big|_{12.5}^{\infty} = \frac{12.5}{20} e^{-20(12.5)} + \frac{1}{400} e^{-20(12.5)}. \end{aligned}$$

Therefore,

$$E(X) = 20e^{20(12.5)} \left[\frac{12.5}{20} e^{-20(12.5)} + \frac{1}{400} e^{-20(12.5)} \right] = 12.5 + \frac{1}{20} = 12.55 \text{ mm.}$$

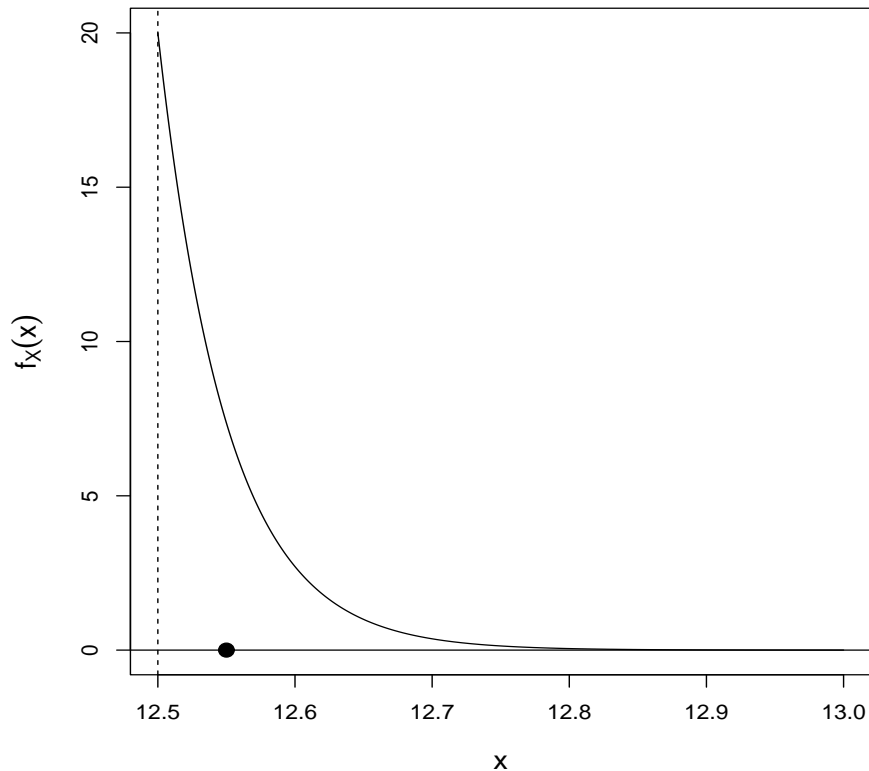


Figure 4.7: Probability density function of X in Example 4.4. The expected value $E(X) = 12.55$ is shown by a solid circle.

R: Here is how to calculate $E(X)$ in R:

```
> x.times.pdf = function(x){x*20*exp(-20*(x-12.5))}
> integrate(x.times.pdf,lower=12.5,upper=Inf) # E(X)
12.55 with absolute error < 1.3e-07
```

Result: Suppose X is a continuous random variable with pdf $f_X(x)$ and g is any function. Then $g(X)$ is also a random variable and its expectation (mean) is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx.$$

Linearity rules: These rules are the same as they were in the discrete case:

- (a) $E(c) = c$, for any constant c
- (b) $E[cg(X)] = cE[g(X)]$, for any constant c
- (c) The expectation of the sum is the sum of the expectations; i.e.,

$$E[g_1(X) + g_2(X) + \cdots + g_k(X)] = E[g_1(X)] + E[g_2(X)] + \cdots + E[g_k(X)].$$

Terminology: Suppose X is a continuous random variable with pdf $f_X(x)$ and mean $\mu = E(X)$. The **variance** of X is

$$\begin{aligned}\sigma^2 = V(X) &= E[(X - \mu)^2] \\ &= \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx.\end{aligned}$$

The **standard deviation** of X is the positive square root of the variance:

$$\sigma = \sqrt{\sigma^2} = \sqrt{V(X)}.$$

Variance computing formula: Suppose X is a random variable (discrete or continuous) with mean $\mu = E(X)$. An alternative way to find $V(X)$ is by using

$$V(X) = E(X^2) - [E(X)]^2.$$

Linear functions: Suppose X is a continuous random variable with pdf $f_X(x)$ and mean $\mu = E(X)$. Suppose a and b are constants. The mean and variance of the linear function $aX + b$ are

$$\begin{aligned}E(aX + b) &= aE(X) + b \\ V(aX + b) &= a^2V(X).\end{aligned}$$

These rules are the same as they were in the discrete case.

Example 4.5. Conductive coatings are applied to a wide variety of materials to make them electrically conductive or to shield them from electromagnetic interference. The thickness of a coating applied to a medical device (measured in micrometers) is a continuous random variable X with pdf

$$f_X(x) = \begin{cases} \frac{1200}{x^2}, & 400 < x < 600 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is shown in Figure 4.8 (next page).

Q: Find $E(X)$ and $V(X)$.

A: The mean is

$$\begin{aligned}E(X) &= \int_{400}^{600} x f_X(x) dx = \int_{400}^{600} \frac{1200x}{x^2} dx \\ &= 1200 \int_{400}^{600} \frac{1}{x} dx \\ &= 1200 \left(\ln x \Big|_{400}^{600} \right) = 1200(\ln 600 - \ln 400) \approx 486.6 \text{ } \mu\text{m}.\end{aligned}$$

To find the variance, let's use the variance computing formula. First, we find

$$E(X^2) = \int_{400}^{600} x^2 f_X(x) dx = \int_{400}^{600} \frac{1200x^2}{x^2} dx = 1200 \int_{400}^{600} 1 dx = 1200(600 - 400) = 240000.$$

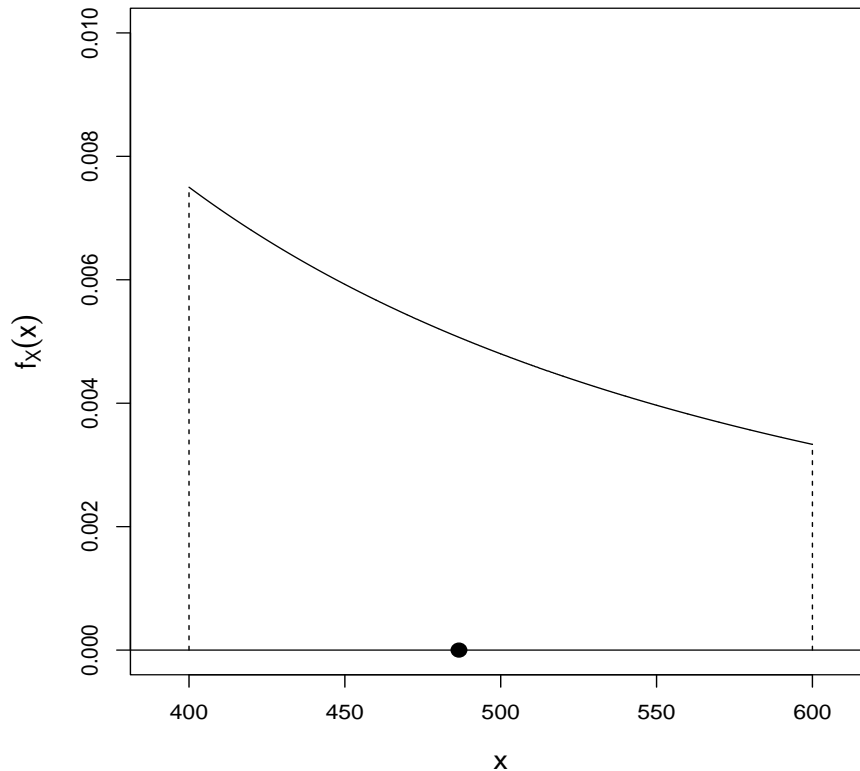


Figure 4.8: Probability density function of X in Example 4.5. The expected value $E(X) \approx 486.6$ is shown by a solid circle.

Therefore,

$$V(X) = E(X^2) - [E(X)]^2 \approx 240000 - (486.6)^2 \approx 3220.4 \text{ } (\mu\text{m})^2.$$

Q: The coating costs $C = 40 + 0.15X$ dollars to apply to each device. That is, there is a fixed cost of 40 dollars plus an additional cost of 0.15 dollars for each micrometer of coating applied. Find $E(C)$ and $V(C)$.

A: The cost C is a linear function of X . We have

$$E(C) = E(40 + 0.15X) = 40 + 0.15E(X) \approx 40 + 0.15(486.6) \approx 112.99 \text{ dollars.}$$

The variance is

$$V(C) = V(40 + 0.15X) = (0.15)^2 V(X) \approx (0.15)^2 (3220.4) \approx 72.46 \text{ (dollars)}^2.$$

The standard deviation of C is

$$\sigma_C = \sqrt{V(C)} \approx \sqrt{72.46} \approx 8.51 \text{ dollars.}$$

Terminology: Suppose X is a **continuous** random variable with cdf $F_X(x)$ and pdf $f_X(x)$. The p th **quantile** ($0 < p < 1$) of X , denoted by ϕ_p , satisfies

$$P(X \leq \phi_p) = \int_{-\infty}^{\phi_p} f_X(x) dx = F_X(\phi_p) = p.$$

In other words, ϕ_p is the value for which 100

percent of the possible values of X are below ϕ_p . Some authors call ϕ_p the 100

th **percentile** of the distribution of X . For example,

- $\phi_{0.10}$ = 10th percentile. This means 10 percent of the values of X are below $\phi_{0.10}$, and 90 percent of the values are above $\phi_{0.10}$.
- $\phi_{0.50}$ = 50th percentile. This means 50 percent of the values of X are below $\phi_{0.50}$, and 50 percent of the values are above $\phi_{0.50}$. This is also called the **median** of X .
- $\phi_{0.99}$ = 99th percentile. This means 99 percent of the values of X are below $\phi_{0.99}$, and 1 percent of the values are above $\phi_{0.99}$.

Example 4.4 (continued). The diameter X (in millimeters) of a hole drilled in a sheet metal component has pdf

$$f_X(x) = \begin{cases} 20e^{-20(x-12.5)}, & x > 12.5 \\ 0, & \text{otherwise.} \end{cases}$$

The cdf of X is

$$F_X(x) = \begin{cases} 0, & x < 12.5 \\ 1 - e^{-20(x-12.5)}, & x > 12.5. \end{cases}$$

The pdf and cdf are shown side by side in Figure 4.9.

Q: Five percent of the diameters will be larger than what value?

A: We want $\phi_{0.95}$, the 95th percentile of the distribution of X . We solve

$$P(X \leq \phi_{0.95}) = \int_{12.5}^{\phi_{0.95}} 20e^{-20(x-12.5)} dx = 1 - e^{-20(\phi_{0.95}-12.5)} = 0.95$$

for $\phi_{0.95}$. This is done as follows:

$$\begin{aligned} 1 - e^{-20(\phi_{0.95}-12.5)} = 0.95 &\implies e^{-20(\phi_{0.95}-12.5)} = 0.05 \\ &\implies -20(\phi_{0.95} - 12.5) = \ln(0.05) \\ &\implies \phi_{0.95} - 12.5 = -\frac{\ln(0.05)}{20} \\ &\implies \phi_{0.95} = 12.5 - \frac{\ln(0.05)}{20} \approx 12.65. \end{aligned}$$

Therefore, approximately 5 percent of the hole diameters will be larger than 12.65 mm.

Exercise: Find $\phi_{0.50}$, the median coating thickness in Example 4.5. How does $\phi_{0.50}$ compare to $E(X)$?

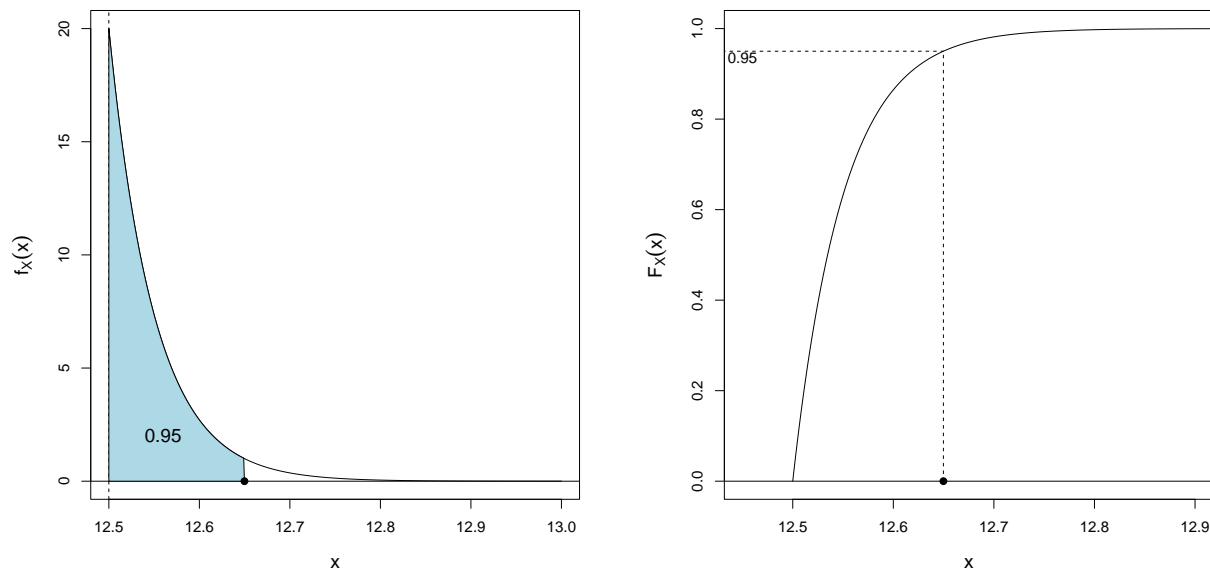


Figure 4.9: Left: Probability density function (pdf) of X in Example 4.4. Right: Cumulative distribution function (cdf) of X . The 95th percentile $\phi_{0.95} \approx 12.65$ is shown in each figure.

4.3 Exponential distribution

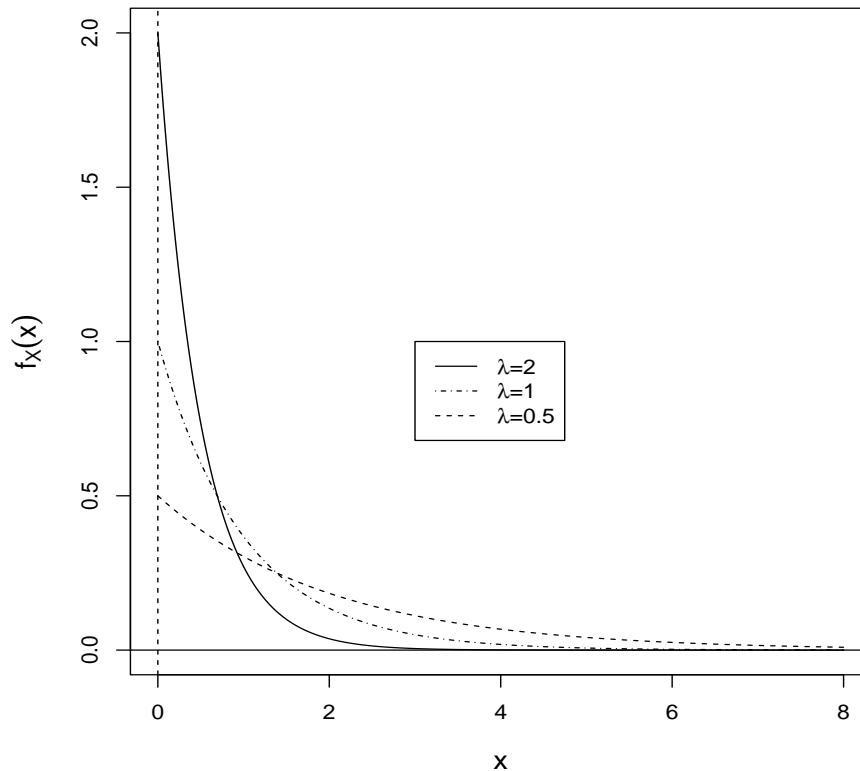
Definition: A continuous random variable X has an **exponential distribution** with parameter $\lambda > 0$ if its pdf is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \text{exponential}(\lambda)$. Different values of λ give different pdfs. All pdfs have the same exponential decay shape; the value of λ controls the scale; see Figure 4.10 (next page).

Remarks:

- The first thing we note is the exponential distribution is for positive quantities ($x > 0$). This includes things like part dimensions, weights, biomarkers, and times.
- The exponential distribution is used in **reliability analysis** and other areas which focus on “time-to-event” random variables, for example,
 - the time until part failure
 - the time until disease onset
 - the time until an insurance claim is filed
 - the time until a catastrophic weather event.

Figure 4.10: Exponential pdfs for different values of λ .

MEAN/VARIANCE: If $X \sim \text{exponential}(\lambda)$, then

$$\begin{aligned} E(X) &= \frac{1}{\lambda} \\ V(X) &= \frac{1}{\lambda^2}. \end{aligned}$$

CDF: If $X \sim \text{exponential}(\lambda)$, then the cdf of X is

$$F_X(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\lambda x}, & x > 0. \end{cases}$$

Example 4.6. The monthly precipitation in Columbia, SC, is a continuous random variable X which is assumed to follow an exponential distribution with mean 4 inches.

Q: What is the probability a given month will have more than 10 inches of precipitation?

A: First note that

$$E(X) = 4 \implies \lambda = 0.25.$$

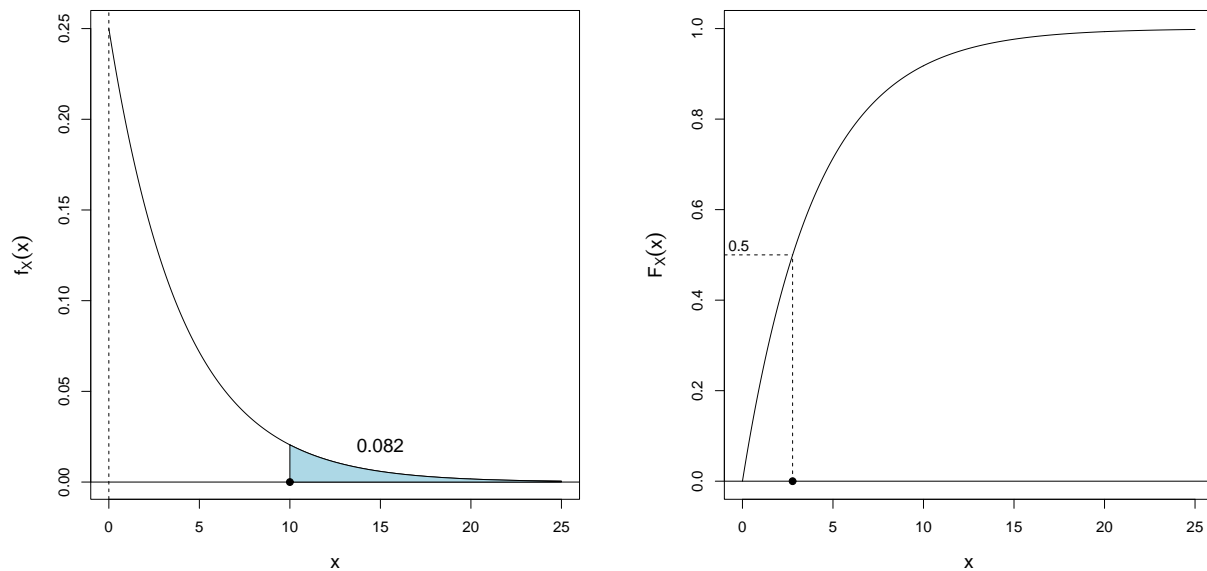


Figure 4.11: Left: Probability density function (pdf) of X in Example 4.6. The probability $P(X > 10) \approx 0.082$ is shown shaded. Right: Cumulative distribution function (cdf) of X . The median $\phi_{0.50} \approx 2.77$ is shown by a solid circle.

The pdf and cdf of X are

$$f_X(x) = \begin{cases} 0.25e^{-0.25x}, & x > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad F_X(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-0.25x}, & x > 0 \end{cases},$$

shown in Figure 4.11 (above). We want

$$\begin{aligned} P(X > 10) &= 1 - P(X \leq 10) \\ &= 1 - F_X(10) \\ &= 1 - [1 - e^{-0.25(10)}] = e^{-2.5} \approx 0.082. \end{aligned}$$

Therefore, approximately 8.2% of the months will have precipitation amounts larger than 10 inches.

Q: What is the median monthly precipitation?

A: We want to solve

$$\begin{aligned} F_X(\phi_{0.50}) = 0.5 &\implies 1 - e^{-0.25\phi_{0.50}} = 0.5 \\ &\implies e^{-0.25\phi_{0.50}} = 0.5 \\ &\implies -0.25\phi_{0.50} = \ln(0.5) \implies \phi_{0.50} = -\frac{\ln(0.5)}{0.25} \approx 2.77 \text{ inches.} \end{aligned}$$

This means 50 percent of the months will have precipitation amounts less than 2.77 inches (and 50 percent of the months will be greater).

EXPONENTIAL R CODE: Suppose $X \sim \text{exponential}(\lambda)$.

$F_X(x) = P(X \leq x)$	ϕ_p
$\text{pexp}(x, \lambda)$	$\text{qexp}(p, \lambda)$

```
> options(digits=3)
> 1-pexp(10,0.25) # 1-P(X<=10)
[1] 0.0821
> qexp(0.5,0.25) # median
[1] 2.77
```

POISSON-EXPONENTIAL RELATIONSHIP: Recall a Poisson distribution arises when we are counting the number of “occurrences” over a unit interval of time (see Section 3.6, notes). Define

$X =$ the **time** until the first occurrence.

This is a continuous random variable and it follows an exponential distribution with parameter λ , where λ is the mean number of occurrences per unit interval of time in the Poisson counting process. That is, $X \sim \text{exponential}(\lambda)$.

Example 4.7. In a corporate computer network, user log-ons to the system are modeled as a Poisson process with a mean of $\lambda = 25$ log-ons per hour.

Q: What is the probability it takes longer than 10 minutes for the first user log-on to occur? Note that 10 minutes = 1/6 of one hour.

A: The time until the first user log-on X follows an exponential distribution with $\lambda = 25$. The pdf and cdf of X are

$$f_X(x) = \begin{cases} 25e^{-25x}, & x > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad F_X(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-25x}, & x > 0 \end{cases}.$$

We want

$$\begin{aligned} P(X > 1/6) &= 1 - P(X \leq 1/6) \\ &= 1 - F_X(1/6) \\ &= 1 - [1 - e^{-25(1/6)}] = e^{-25/6} \approx 0.016. \end{aligned}$$

```
> 1-pexp(1/6,25) # 1-P(X<=1/6)
[1] 0.016
```

Remark: Another interesting fact is the time between any two successive occurrences in a Poisson process follows the same $\text{exponential}(\lambda)$ distribution. Times between successive occurrences are called **interarrival times**. In Example 4.7,

- the time until the first user log-on is $\text{exponential}(\lambda = 25)$,
- the time between the first user log-on and the second is $\text{exponential}(\lambda = 25)$,
- the time between the second user log-on and the third is $\text{exponential}(\lambda = 25)$,

and so on.

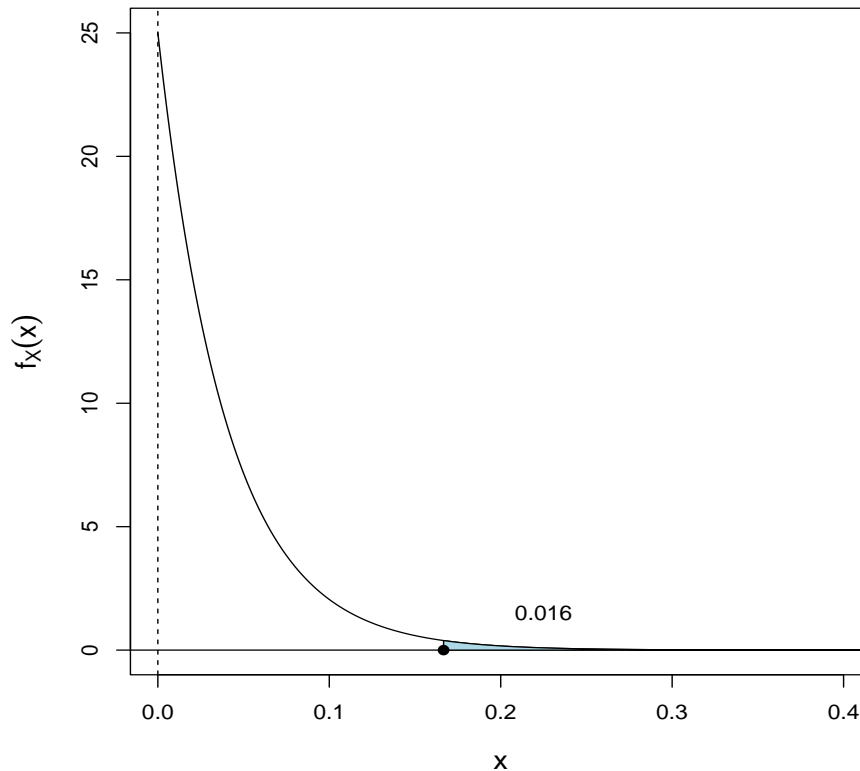


Figure 4.12: Probability density function of X in Example 4.7. The shaded area under the curve is $P(X > 1/6) \approx 0.016$.

MEMORYLESS PROPERTY: A unique property of the exponential distribution is its “lack of memory.” Suppose a continuous random variable X measures the time until some event occurs (e.g., part failure, disease onset, claim is filed, catastrophic weather, etc.). If $X \sim \text{exponential}(\lambda)$, then

$$P(X > t_1 + t_2 | X > t_1) = P(X > t_2).$$

In the context of a part failing, here is how this can be interpreted:

- We have one part in the field whose failure time X is known to be larger than t_1 , that is, the part has been in operation and has not failed before time t_1 .
- We have a second part that has just been put in operation (at “time zero”).
- The memoryless property says the probability the first part does not fail before an additional time of t_2 is the same as the second part not failing before time t_2 . In other words, the fact the first part has been in operation for time t_1 has been “forgotten.”

Remark: The memoryless property is mandated whenever one makes an exponential distribution assumption about the time to event. It is a restrictive condition and may not be realistic. For example, in the part failure context, an exponential distribution assumption for X requires that parts in the field do not “wear out” or “get stronger” over time.

Example 4.8. At a hospital’s intensive care unit (ICU), the time until patient discharge (in days) is a continuous random variable X which is assumed to have an exponential distribution with $\lambda = 1/7$. The pdf and cdf of X are

$$f_X(x) = \begin{cases} \frac{1}{7}e^{-x/7}, & x > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad F_X(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-x/7}, & x > 0 \end{cases}.$$

Q: One patient has been in the ICU for 5 days. What is the probability the patient is still in the ICU after 8 days?

A: From the memoryless property, we know

$$P(X > 8 | X > 5) = P(X > 3).$$

This equals

$$1 - P(X \leq 3) = 1 - F_X(3) = e^{-3/7} \approx 0.65.$$

4.4 Gamma distribution

Definition: A continuous random variable X has a **gamma distribution** with parameters $r > 0$ and $\lambda > 0$ if its pdf is given by

$$f_X(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \text{gamma}(r, \lambda)$. We call

- r = shape parameter
- λ = scale parameter.

Different values of r and λ give different pdfs; see Figure 4.13 (next page). The gamma distribution is more flexible than the exponential distribution. Introducing the extra parameter r allows for different shapes, whereas the exponential distribution imposes the same exponential decay shape regardless of what λ is.

Q: What is $\Gamma(r)$?

A: It is a constant defined as the following integral

$$\Gamma(r) = \int_0^\infty u^{r-1} e^{-u} du,$$

provided that $r > 0$. In mathematical analysis, this is called the **gamma function**.

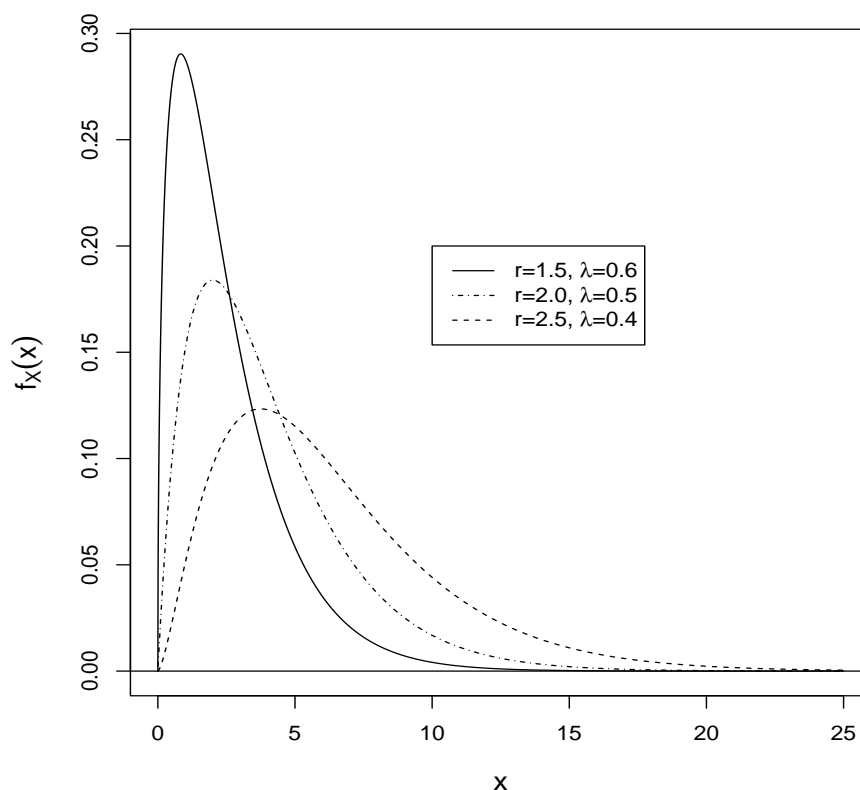


Figure 4.13: Gamma pdfs for different values of r and λ .

Important: When $r = 1$, the $\text{gamma}(r, \lambda)$ distribution reduces to the $\text{exponential}(\lambda)$ distribution. This is true because

$$\Gamma(1) = \int_0^{\infty} e^{-u} du = 1$$

and therefore the $\text{gamma}(r, \lambda)$ pdf

$$\frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} = \lambda e^{-\lambda x}, \quad \text{provided that } r = 1.$$

GAMMA R CODE: Suppose $X \sim \text{gamma}(r, \lambda)$.

$F_X(x) = P(X \leq x)$	ϕ_p
<code>pgamma(x,r,λ)</code>	<code>qgamma(p,r,λ)</code>

Note: Probability and quantile calculations for the gamma distribution can be carried out using R. The cdf of $X \sim \text{gamma}(r, \lambda)$ does not exist in closed form for all values of $r > 0$, so numerical evaluation is required.

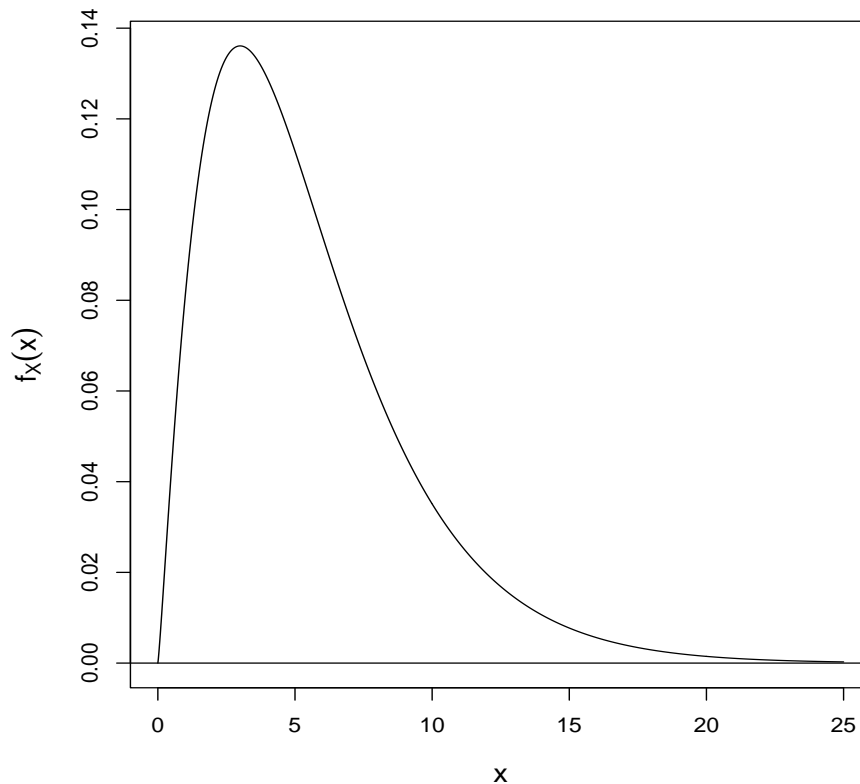


Figure 4.14: Probability density function of X in Example 4.9.

Example 4.9. The lifetime of a diesel engine fan blade X (in 1000s of hours) is modeled using a gamma distribution with $r = 2.2$ and $\lambda = 0.4$. This pdf is shown in Figure 4.14.

Q: What is the probability a fan blade fails before 5000 hours of operation?

A: We want

$$P(X < 5) = \int_0^5 \frac{(0.4)^{2.2}}{\Gamma(2.2)} x^{1.2} e^{-0.4x} dx \approx 0.536.$$

This probability is calculated numerically using the R code:

```
> options(digits=3)
> pgamma(5,2.2,0.4)
[1] 0.536
```

Q: Find the 90th percentile of this distribution and interpret what it means.

```
> options(digits=5)
> qgamma(0.9,2.2,0.4)
[1] 10.461
```

Ninety percent of all fan blades will fail before 10,461 hours of operation.

MEAN/VARIANCE: If $X \sim \text{gamma}(r, \lambda)$, then

$$\begin{aligned} E(X) &= \frac{r}{\lambda} \\ V(X) &= \frac{r}{\lambda^2}. \end{aligned}$$

Letting $r = 1$ in the formulas above gives $E(X)$ and $V(X)$ for $X \sim \text{exponential}(\lambda)$.

POISSON-GAMMA RELATIONSHIP: Recall a Poisson distribution arises when we are counting the number of “occurrences” over a unit interval of time (see Section 3.6, notes). Define

$X =$ the **time** until the r th occurrence.

This is a continuous random variable and it follows a $\text{gamma}(r, \lambda)$ distribution, where λ is the mean number of occurrences per unit interval of time in the Poisson counting process. That is, $X \sim \text{gamma}(r, \lambda)$.

- Of course, if $r = 1$, then X is the time until the **first** occurrence, which we know is $\text{exponential}(\lambda)$.

Example 4.7 (continued). In a corporate computer network, user log-ons to the system are modeled as a Poisson process with a mean of $\lambda = 25$ log-ons per hour.

- the time until the first user log-on is $\text{exponential}(\lambda = 25)$,
- the time until the second user log-on is $\text{gamma}(r = 2, \lambda = 25)$,
- the time until the third user log-on is $\text{gamma}(r = 3, \lambda = 25)$,

and so on.

4.5 Normal distribution

Definition: A continuous random variable X has a **normal distribution** with mean μ and variance σ^2 if its pdf is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad \text{for } -\infty < x < \infty.$$

We write $X \sim \mathcal{N}(\mu, \sigma^2)$. This is also called the **Guassian distribution**. The parameters μ and σ^2 are the mean and variance of X , respectively, that is,

$$\begin{aligned} E(X) &= \mu \\ V(X) &= \sigma^2. \end{aligned}$$

The mean μ identifies where the “center” of the distribution is. The variance σ^2 measures the “spread” of the distribution.

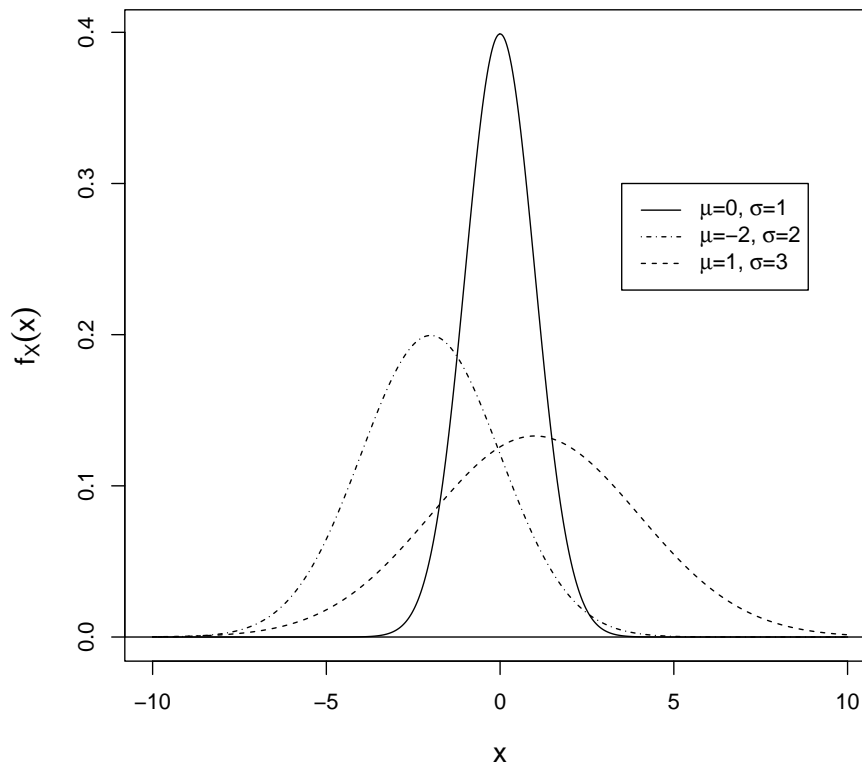


Figure 4.15: Normal pdfs for different values of μ and σ^2 . Note the standard deviation σ is used instead of the variance σ^2 .

Facts:

- The $\mathcal{N}(\mu, \sigma^2)$ pdf is **symmetric** about the mean μ .
- The $\mathcal{N}(\mu, \sigma^2)$ pdf has points of inflection at $\mu - \sigma$ and $\mu + \sigma$.
- The $\mathcal{N}(\mu, \sigma^2)$ pdf follows the **68-95-99.7% Rule**:

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &\approx 0.68 \\ P(\mu - 2\sigma < X < \mu + 2\sigma) &\approx 0.95 \\ P(\mu - 3\sigma < X < \mu + 3\sigma) &\approx 0.997. \end{aligned}$$

That is, approximately 68% of the observations x will be within 1 standard deviation of the mean, approximately 95% of the observations will be within 2 standard deviations of the mean, and approximately 99.7% of the observations will be within 3 standard deviations of the mean.

- It is unlikely for a $\mathcal{N}(\mu, \sigma^2)$ random variable to have a value x further than 3 standard deviations away from its mean in either direction; this probability is approximately 0.003 or 0.3%.

NORMAL R CODE: Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$.

$F_X(x) = P(X \leq x)$	ϕ_p
<code>pnorm(x, μ, σ)</code>	<code>qnorm(p, μ, σ)</code>

Note: Probability and quantile calculations for the normal distribution can be carried out using R. The cdf of $X \sim \mathcal{N}(\mu, \sigma^2)$ does not exist in closed form, so numerical evaluation is required. Note that R parameterizes the normal distribution by the standard deviation σ (not the variance σ^2).

Example 4.10. The daily demand for water use in Atlanta, GA, is a continuous random variable X , measured in millions of gallons. Suppose the distribution of X is normal (Gaussian) with mean $\mu = 447$ and standard deviation $\sigma = 32$. This pdf is shown in Figure 4.16 (next page).

Q: What is the probability the daily water demand will be less than 400 million gallons?

A: We want

$$P(X < 400) = \int_{-\infty}^{400} \frac{1}{\sqrt{2\pi}(32)} e^{-(x-447)^2/2(32)^2} dx \approx 0.071.$$

This probability is calculated numerically using the R code:

```
> options(digits=3)
> pnorm(400,447,32)
[1] 0.071
```

Q: City reservoirs are filled daily to a designated capacity. What capacity is needed so that the probability the daily demand exceeds the capacity is only 0.01?

A: We want $\phi_{0.99}$, the 99th percentile of this distribution. From R,

```
> options(digits=9)
> qnorm(0.99,447,32)
[1] 521.443132
```

Therefore, the capacity should be set at 521,443,132 gallons.

Terminology: A normal random variable with mean 0 and variance 1 is called a **standard normal** random variable. The pdf of $Z \sim \mathcal{N}(0, 1)$ is

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \text{for } -\infty < z < \infty.$$

The cdf of Z can be written as

$$F_Z(z) = P(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

The term $e^{-t^2/2}$ in the integrand does not have an antiderivative in closed form, so probabilities and quantiles associated with the standard normal distribution (and, in fact, any normal distribution) must be calculated numerically.

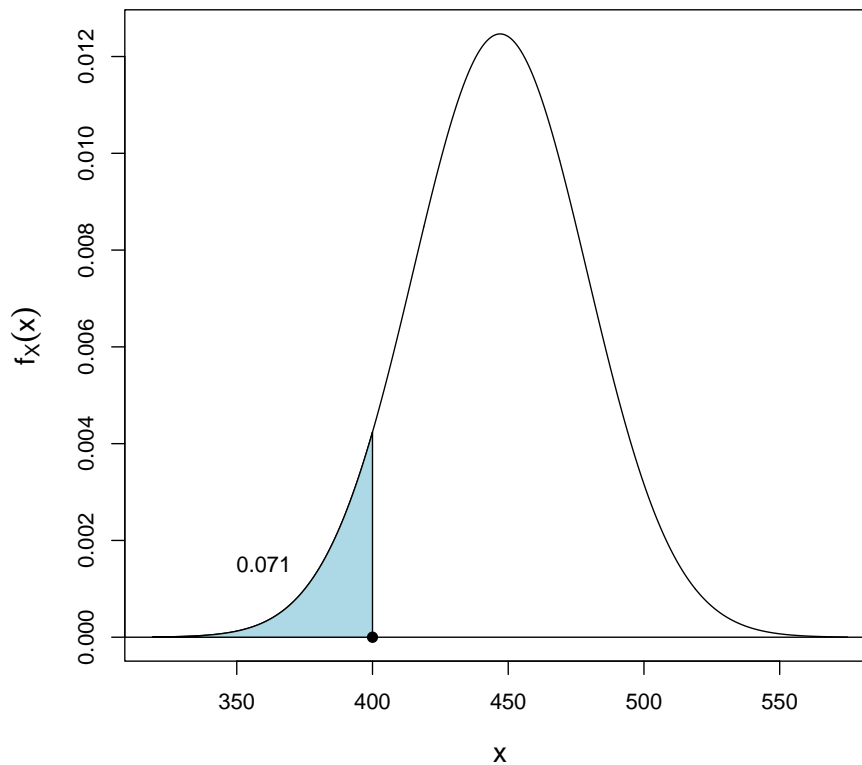


Figure 4.16: Probability density function of X in Example 4.10. The shaded area under the curve is $P(X < 400) \approx 0.071$.

Important: If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

This result says any normal random variable X can be “converted” to a standard normal random variable Z by applying this linear transformation. This conversion is known as **standardization**. For example, suppose exam scores for a population of students have mean $\mu = 70$ and standard deviation $\sigma = 10$. A student’s score of 85 has standardized value

$$z = \frac{85 - 70}{10} = 1.5.$$

This means the student’s score is 1.5 standard deviations above the mean. Similarly, a student’s score of 55 produces

$$z = \frac{55 - 70}{10} = -1.5,$$

meaning the score is 1.5 standard deviations below the mean. From the 68-95-99.7% Rule, we know almost all standardized values z will be between -3 and 3 .

5 Reliability Analysis and Lifetime Distributions

Terminology: Reliability analysis deals with the analysis of “time-to-event data.” This means we are interested in a continuous random variable T which measures the time until something occurs. For example,

$$\begin{aligned} T &= \text{time until part failure} \\ T &= \text{time until maintenance is required} \\ T &= \text{time until a warranty claim is filed} \\ T &= \text{lifespan of a biological organism.} \end{aligned}$$

It is understood we are measuring something for which there is an unambiguous start and end, with the time in between corresponding to T . We call T a **lifetime random variable** because $P(T \geq 0) = 1$, that is, T assumes positive values only.

Terminology: A lifetime distribution describes the distribution of a lifetime random variable T . It has positive support. Some common choices are

- Weibull ← by far the most common in engineering applications
- lognormal
- gamma
- exponential (arises as a special gamma and as a special Weibull).

Although the normal distribution is the most widely used distribution in all of statistics, it is rarely used for reliability analyses. Typical time-to-event data are positive and skewed to the right. These characteristics are incongruous with normal distributions.

5.1 Weibull distribution

Definition: A continuous random variable T has a **Weibull distribution** with parameters $\beta > 0$ and $\eta > 0$ if its pdf is given by

$$f_T(t) = \begin{cases} \frac{\beta}{\eta} \left(\frac{t}{\eta}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\eta}\right)^\beta\right], & t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We write $T \sim \text{Weibull}(\beta, \eta)$. We call

- β = shape parameter
- η = scale parameter.

Different values of β and η give different pdfs; see Figure 5.1 (next page). Weibull distributions have positive support and are generally skewed to the right.

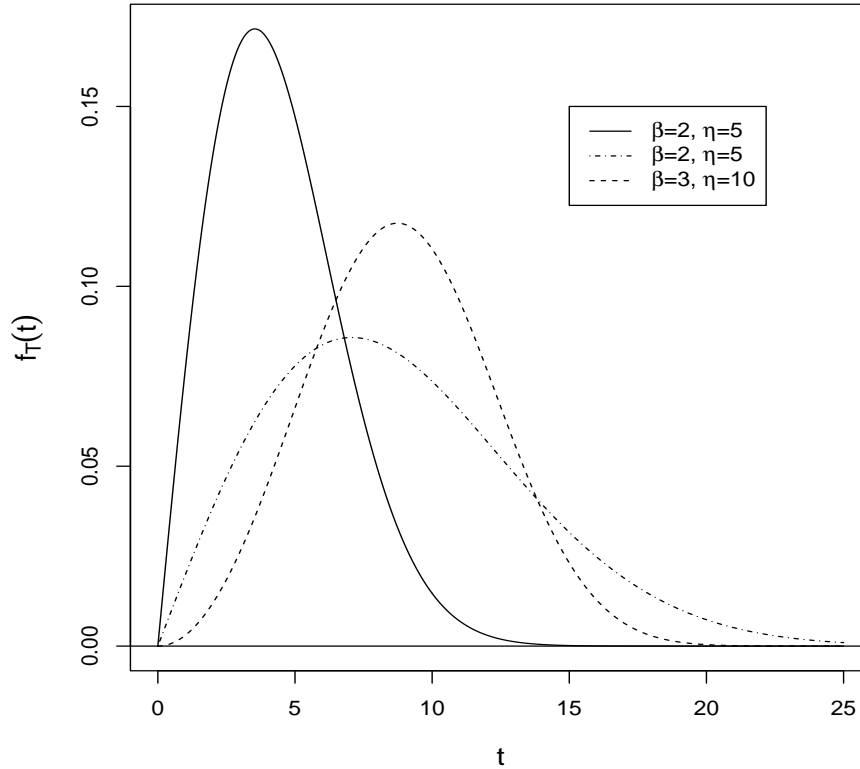


Figure 5.1: Weibull pdfs for different values of β and η .

Important: When $\beta = 1$, the $\text{Weibull}(\beta, \eta)$ distribution reduces to the $\text{exponential}(\lambda)$ distribution, where $\lambda = 1/\eta$. This is true because the $\text{Weibull}(\beta, \eta)$ pdf

$$\frac{\beta}{\eta} \left(\frac{t}{\eta} \right)^{\beta-1} \exp \left[- \left(\frac{t}{\eta} \right)^{\beta} \right] = \frac{1}{\eta} e^{-t/\eta}, \quad \text{provided that } \beta = 1.$$

MEAN/VARIANCE: If $T \sim \text{Weibull}(\beta, \eta)$, then

$$\begin{aligned} E(T) &= \eta \Gamma \left(1 + \frac{1}{\beta} \right) \\ V(T) &= \eta^2 \left\{ \Gamma \left(1 + \frac{2}{\beta} \right) - \left[\Gamma \left(1 + \frac{1}{\beta} \right) \right]^2 \right\}, \end{aligned}$$

where recall $\Gamma(\cdot)$ is the gamma function defined in Section 4.4 (notes). The R code `gamma(r)` will calculate

$$\Gamma(r) = \int_0^{\infty} u^{r-1} e^{-u} du,$$

for any $r > 0$.

CDF: If $T \sim \text{Weibull}(\beta, \eta)$, then the cdf of T is

$$F_T(t) = \begin{cases} 0, & t \leq 0 \\ 1 - \exp \left[- \left(\frac{t}{\eta} \right)^\beta \right], & t > 0. \end{cases}$$

Therefore, Weibull probabilities of the form $F_T(b) = P(T \leq b)$ and

$$P(a < T < b) = F_T(b) - F_T(a)$$

can be calculated without using numerical methods. Quantiles ϕ_p can be found by solving

$$F_T(\phi_p) = 1 - \exp \left[- \left(\frac{\phi_p}{\eta} \right)^\beta \right] = p.$$

WEIBULL R CODE: Suppose $T \sim \text{Weibull}(\beta, \eta)$.

$F_T(t) = P(T \leq t)$	ϕ_p
<code>pweibull(t, beta, eta)</code>	<code>qweibull(p, beta, eta)</code>

Example 5.1. In a mechanical assembly, a bearing allows a shaft to rotate smoothly with minimal friction. The time until the bearing fails (in hours) is modeled as a Weibull random variable T with $\beta = 1.5$ and $\eta = 3000$. The pdf and cdf of T are shown side by side in Figure 5.2 (next page).

Q: What is the probability the bearing fails before 5000 hours of operation?

A: We want

$$P(T < 5000) = F_T(5000) = 1 - \exp \left[- \left(\frac{5000}{3000} \right)^{1.5} \right] = 1 - e^{-(5/3)^{1.5}} \approx 0.884.$$

That is, 88.4% of all bearings will fail before 5000 hours of operation.

```
> options(digits=3)
> pweibull(5000,1.5,3000)
[1] 0.884
```

Remark: When T measures the time to failure, we can think of the cdf $F_T(t)$ as the proportion of all units (here, shaft bearings) in the population which have “failed” before time t . Of course, if a unit has not failed before time t , then it is still operational and hence has “survived” up until time t . We call

$$S_T(t) = 1 - F_T(t)$$

the **survivor function** for this reason. It represents the proportion of all units in the population still “alive” at time t .

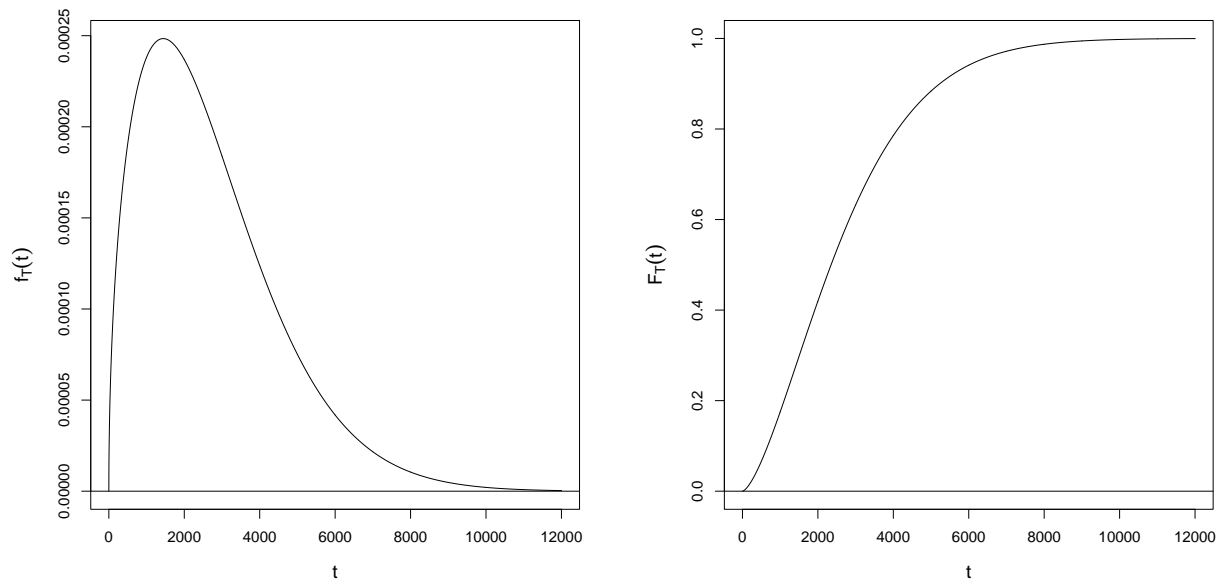


Figure 5.2: Left: Probability density function (pdf) of T in Example 5.1. Right: Cumulative distribution function (cdf) of T .

Q: Find the mean and median time to bearing failure.

A: The mean time to failure is

$$E(T) = 3000 \Gamma\left(1 + \frac{1}{1.5}\right) = 3000 \Gamma\left(\frac{5}{3}\right) \approx 2708.2 \text{ hours.}$$

The median time to failure $\phi_{0.5}$ solves

$$\begin{aligned} F_T(\phi_{0.5}) = 1 - \exp\left[-\left(\frac{\phi_{0.5}}{3000}\right)^{1.5}\right] &= 0.5 \implies \exp\left[-\left(\frac{\phi_{0.5}}{3000}\right)^{1.5}\right] = 0.5 \\ \implies -\left(\frac{\phi_{0.5}}{3000}\right)^{1.5} &= \ln(0.5) \\ \implies \left(\frac{\phi_{0.5}}{3000}\right)^{1.5} &= -\ln(0.5) \\ \implies \frac{\phi_{0.5}}{3000} &= [-\ln(0.5)]^{1/1.5} \\ \implies \phi_{0.5} &= 3000[-\ln(0.5)]^{1/1.5} \approx 2349.7 \text{ hours.} \end{aligned}$$

```
> options(digits=5)
> 3000*gamma(5/3) # E(T)
[1] 2708.2
> qweibull(0.5,1.5,3000) # median
[1] 2349.7
```

5.2 Reliability functions

Goal: We now summarize some different, but equivalent, ways of defining the distribution of a continuous lifetime random variable T . We also introduce a new function which is used in reliability studies.

- The **cumulative distribution function** (cdf)

$$F_T(t) = P(T \leq t).$$

This can be interpreted as the proportion of units (individuals) in the population that have failed at or before time t .

- The **survivor function**

$$S_T(t) = P(T > t) = 1 - F_T(t).$$

This can be interpreted as the proportion of units (individuals) in the population that have not failed by time t ; e.g., unit is still operational, warranty claim has not been made, organism is still alive, etc.

- The **probability density function** (pdf)

$$f_T(t) = \frac{d}{dt}F_T(t) = -\frac{d}{dt}S_T(t).$$

Also, recall

$$F_T(t) = \int_0^t f_T(u)du \longleftarrow \text{area under the pdf over } (0, t)$$

and

$$S_T(t) = \int_t^\infty f_T(u)du \longleftarrow \text{area under the pdf over } (t, \infty).$$

Terminology: The **hazard function** of a lifetime random variable T is defined as

$$h_T(t) = \lim_{\epsilon \rightarrow 0} \frac{P(t \leq T < t + \epsilon | T \geq t)}{\epsilon},$$

for $\epsilon > 0$. The hazard function is not a probability; rather, it is a **probability rate**. It characterizes the instantaneous potential for failure to occur, given that a unit (individual) has already survived up to a certain point in time t .

Interpretation: The hazard function offers a useful interpretation. *It indicates how the rate of failure varies with time.*

- Distributions with increasing hazard functions are seen in units (individuals) where some kind of aging or “wear out” takes place. The population gets weaker over time.

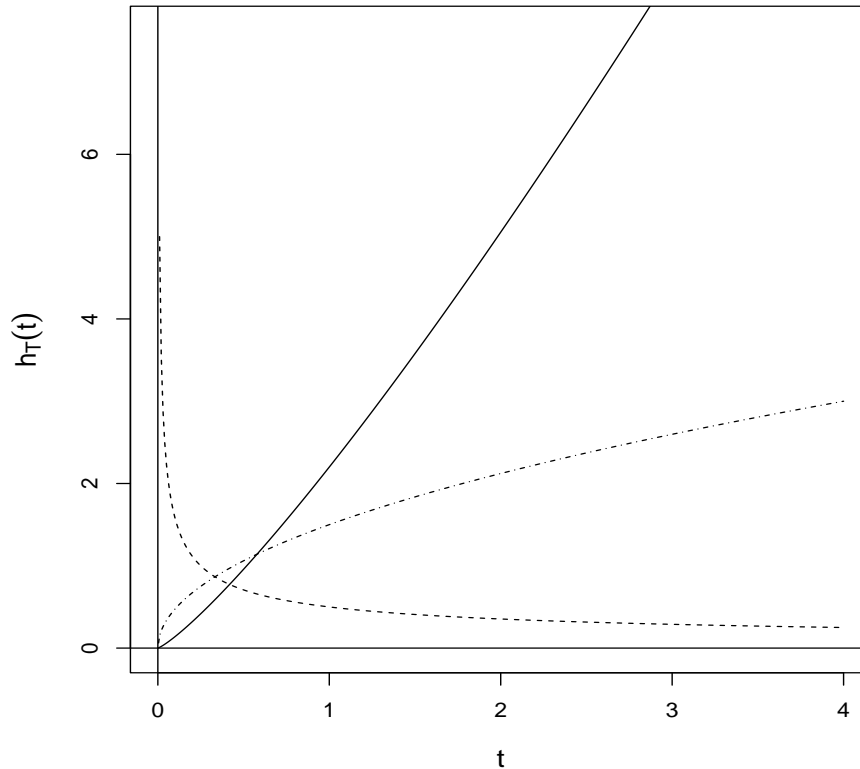


Figure 5.3: Examples of hazard functions. Increasing hazards correspond to the population getting weaker over time.

- Distributions with decreasing hazard functions correspond to a population getting stronger over time. This is observed in scenarios like the “infant mortality” phase in manufacturing, where initial defects are weeded out, or in situations where components become more robust with use.
- In some populations, the hazard function decreases initially, stays constant for a period of time, and then increases. This corresponds to a population whose units get stronger initially (defective units “die out” early), exhibit random failures for a period of time (constant hazard), and then eventually the population starts to weaken (e.g., due to wear/old age, etc.). These hazard functions are **bathtub-shaped**.

Result: Suppose T is a lifetime random variable with pdf $f_T(t)$ and survivor function $S_T(t)$. The hazard function

$$h_T(t) = \frac{f_T(t)}{S_T(t)}.$$

We can therefore describe the distribution of T by using either $f_T(t)$, $F_T(t)$, $S_T(t)$, or $h_T(t)$. If we know one of these functions, we can always retrieve the other three.

Note: The previous result can be shown by using the definitions of conditional probability (Chapter 2) and of the derivative from calculus. Note that

$$\begin{aligned}
 h_T(t) &= \lim_{\epsilon \rightarrow 0} \frac{P(t \leq T < t + \epsilon | T \geq t)}{\epsilon} \\
 &= \lim_{\epsilon \rightarrow 0} \frac{P(t \leq T < t + \epsilon \text{ and } T \geq t)}{\epsilon P(T \geq t)} \\
 &= \lim_{\epsilon \rightarrow 0} \frac{P(t \leq T < t + \epsilon)}{\epsilon S_T(t)} \\
 &= \frac{1}{S_T(t)} \underbrace{\lim_{\epsilon \rightarrow 0} \frac{F_T(t + \epsilon) - F_T(t)}{\epsilon}}_{= \frac{d}{dt} F_T(t)} = \frac{f_T(t)}{S_T(t)}.
 \end{aligned}$$

WEIBULL HAZARD: If $T \sim \text{Weibull}(\eta, \beta)$, then the hazard function of T is

$$h_T(t) = \frac{f_T(t)}{S_T(t)} = \frac{f_T(t)}{1 - F_T(t)} = \frac{\frac{\beta}{\eta} \left(\frac{t}{\eta}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\eta}\right)^\beta\right]}{1 - \left\{1 - \exp\left[-\left(\frac{t}{\eta}\right)^\beta\right]\right\}} = \frac{\beta}{\eta} \left(\frac{t}{\eta}\right)^{\beta-1}.$$

Interpretation: For a Weibull distribution,

- $h_T(t)$ is **increasing** if $\beta > 1$ (wear out; population gets weaker)
- $h_T(t)$ is **constant** if $\beta = 1$ (random failures; exponential distribution)
- $h_T(t)$ is **decreasing** if $\beta < 1$ (infant mortality; population gets stronger).

In other words, the value of the shape parameter β completely summarizes the relevant feature of the hazard function. This is one reason the Weibull distribution is popular. Engineers can characterize the rate of failure over time by knowing this single number. Other lifetime distributions (e.g., gamma, lognormal, etc.) have hazard functions which are more complex. They are not amenable to this easy interpretation.

Example 5.1 (continued). In a mechanical assembly, a bearing allows a shaft to rotate smoothly with minimal friction. The time until the bearing fails (in hours) is modeled as a Weibull random variable T with $\beta = 1.5$ and $\eta = 3000$. The hazard function

$$h_T(t) = \frac{1.5}{3000} \left(\frac{t}{3000}\right)^{1.5-1} = \frac{1.5}{(3000)^{3/2}} t^{1/2}$$

is an increasing function of t ; see Figure 5.4 (next page). This means that (under the Weibull model assumption) the rate of bearing failure increases over time. This corresponds to “aging” or “wear out” in the population of bearings over time.

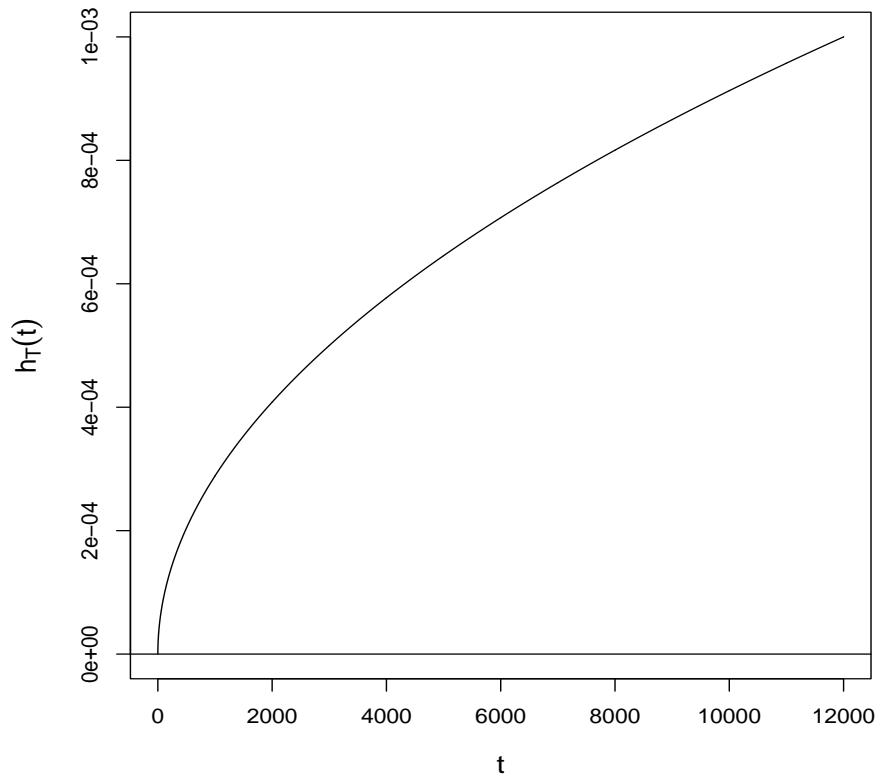


Figure 5.4: Hazard function of T in Example 5.1.

5.3 Fitting a Weibull distribution to data

Curiosity: In Example 5.1, we used a Weibull distribution with $\beta = 1.5$ and $\eta = 3000$ to model T , the time until shaft bearing failure. A natural question to ask is, “Where do the values of η and β come from?” or “How do we know these values are correct?”

- Formulating a good answer to the first question is easy, sort of. The reason for the ambiguity is that β and η are **population parameters**. They represent the shape and scale of the Weibull distribution that is selected to model the time to failure for all shaft bearings in the population.
- Therefore, the only way we could determine β and η without ambiguity is to observe the failure time T for all shaft bearings in the population! This is not possible. For this reason, population parameters like η and β will be unknown in real life (and, thus, no one can answer the second question above).
- What can we do? We do the next best thing. If we observe a **sample** of shaft bearings from the population, we can find **estimates** of η and β by using the failure times in

the sample. If the sample is representative of the population, then the estimates of η and β we calculate should be reasonable “guesses” of what the true η and β are.

- Of course, a deeper foundational question is “How do we know if the Weibull distribution is the correct lifetime distribution for the population of all shaft bearings?” Again, this question cannot be answered, because “correct” is too strong a word (it would require us to observe the failure time of all shaft bearings in the population). We *can* assess empirically whether the Weibull model is “reasonable” by using a representative sample. This is done in the next section.

Example 5.2. A shock absorber is a suspension component that controls the up-and-down motion of a vehicle’s wheels. The following data are $n = 38$ distances (in km) driven to failure for a specific brand of shock absorber under extreme driving conditions.

6700	6950	7820	9120	9660	9820	11310	11690	11850	11880
12140	12200	12870	13150	13330	13470	14040	14300	17520	17540
17890	18450	18960	18980	19410	20100	20100	20150	20320	20900
22700	23490	26510	27410	27490	27890	28100	30050		

We will assume a Weibull(β, η) distribution for

$$T = \text{distance until failure (in km).}$$

Because the population parameters β and η are not given to us, our first task is to estimate them. We do this by finding the values of β and η that “most closely agree” with the data above. Form the **likelihood function**

$$\begin{aligned} L(\beta, \eta) = \prod_{i=1}^{38} f_T(t_i) &= \prod_{i=1}^{38} \frac{\beta}{\eta} \left(\frac{t_i}{\eta} \right)^{\beta-1} \exp \left[- \left(\frac{t_i}{\eta} \right)^{\beta} \right] \\ &= \left(\frac{\beta}{\eta^{\beta}} \right)^{38} \left(\prod_{i=1}^{38} t_i \right)^{\beta-1} \exp \left[- \sum_{i=1}^{38} \left(\frac{t_i}{\eta} \right)^{\beta} \right], \end{aligned}$$

where t_1, t_2, \dots, t_{38} are the 38 distances. Informally, the likelihood function describes the probability of the observed data. Therefore, the values of β and η that “most closely agree” with the data are the values that maximize $L(\beta, \eta)$.

- Let $\hat{\beta}$ and $\hat{\eta}$ denote the values of β and η , respectively, that maximize $L(\beta, \eta)$. We call $\hat{\beta}$ and $\hat{\eta}$ **maximum likelihood estimates**.
- Finding $\hat{\beta}$ and $\hat{\eta}$ is a multivariable calculus problem we will solve numerically using R.
- In statistics speak, we say that $\hat{\beta}$ and $\hat{\eta}$ are **estimates** of the population parameters β and η , respectively. The population here is the universe of all shock absorbers of this specific brand.

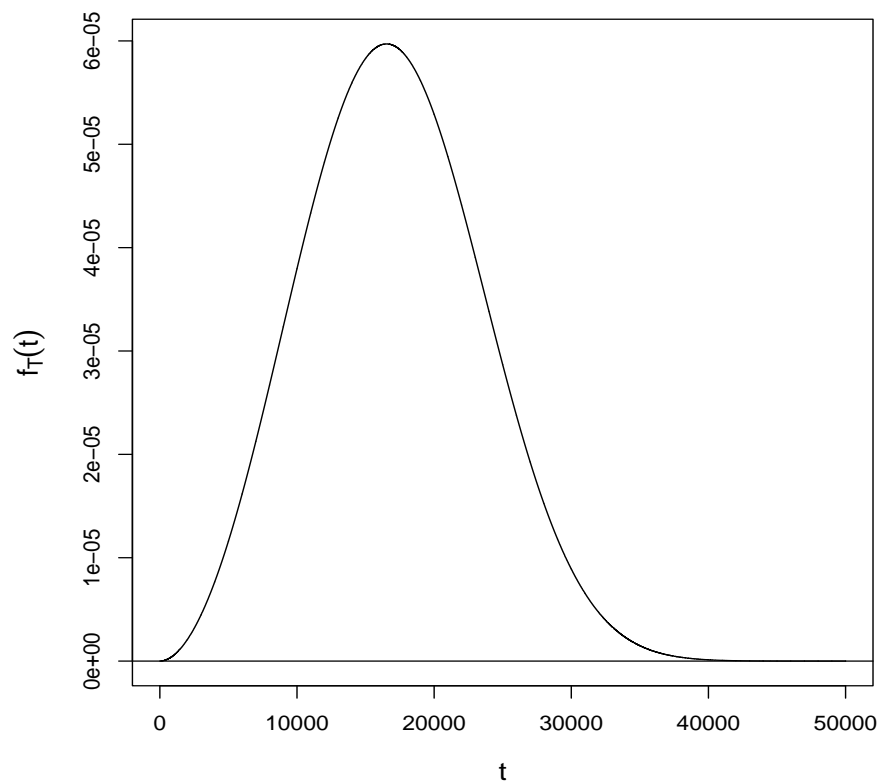


Figure 5.5: (Estimated) probability density function of T in Example 5.2.

Implementation in R: We can use the `fitdistrplus` package:

```
> library(fitdistrplus)
> distance.to.failure = c(6700,6950,7820, ..., 30050) # Enter the data
> options(digits=3)
> fitdist(distance.to.failure,distr="weibull",method="mle")
```

Fitting of the distribution ' weibull ' by maximum likelihood

Parameters:

	estimate	Std. Error
shape	2.9	0.367
scale	19125.6	1140.512

This output produces

$$\begin{aligned}\hat{\beta} &\approx 2.9 \\ \hat{\eta} &\approx 19125.6.\end{aligned}$$

These are the **estimates** of β and η based on the data from the previous page.

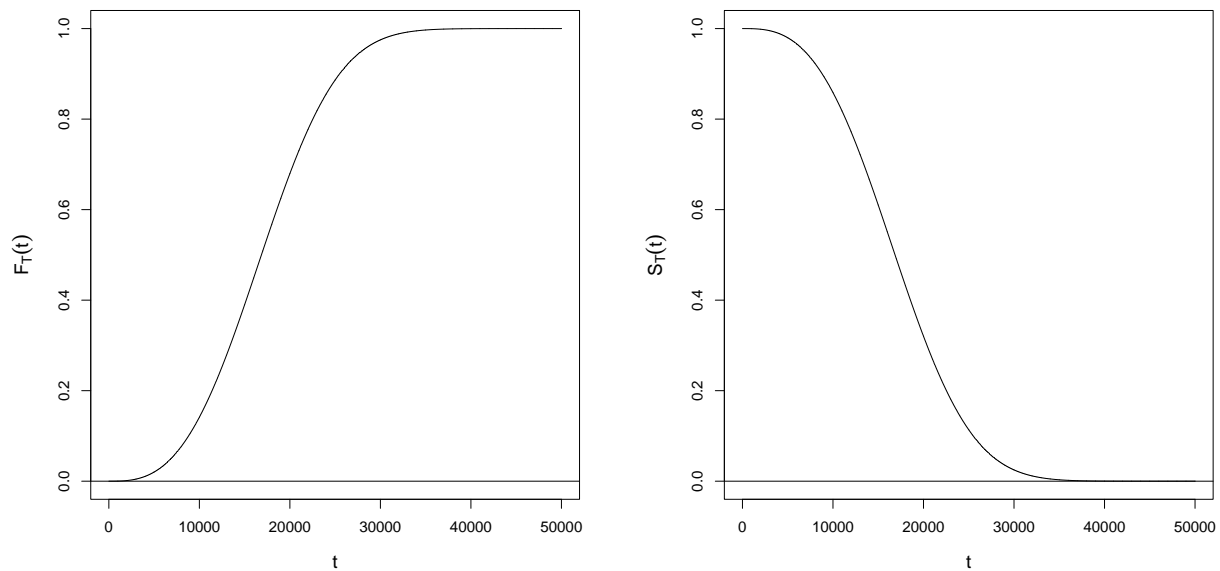


Figure 5.6: Left: (Estimated) cumulative distribution function T in Example 5.2. Right: (Estimated) survivor function of T .

Here are the estimated functions for the shock absorber data in Example 5.2:

PDF:

$$f_T(t) = \begin{cases} \frac{2.9}{19125.6} \left(\frac{t}{19125.6} \right)^{1.9} \exp \left[- \left(\frac{t}{19125.6} \right)^{2.9} \right], & t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

CDF/Survivor:

$$F_T(t) = \begin{cases} 0, & t \leq 0 \\ 1 - \exp \left[- \left(\frac{t}{19125.6} \right)^{2.9} \right], & t > 0 \end{cases} \quad S_T(t) = \begin{cases} 1, & t \leq 0 \\ \exp \left[- \left(\frac{t}{19125.6} \right)^{2.9} \right], & t > 0 \end{cases}.$$

Q: Estimate the proportion of shock absorbers in this population that will still be operational at 30,000 km.

A: We want

$$P(T \geq 30000) = S_T(30000) = \exp \left[- \left(\frac{30000}{19125.6} \right)^{2.9} \right] \approx 0.025.$$

About 2.5% of all shock absorbers in the population will still be operational at 30,000 km.

```
> 1-pweibull(30000,2.9,19125.6) # P(T>=30000)
[1] 0.025
```

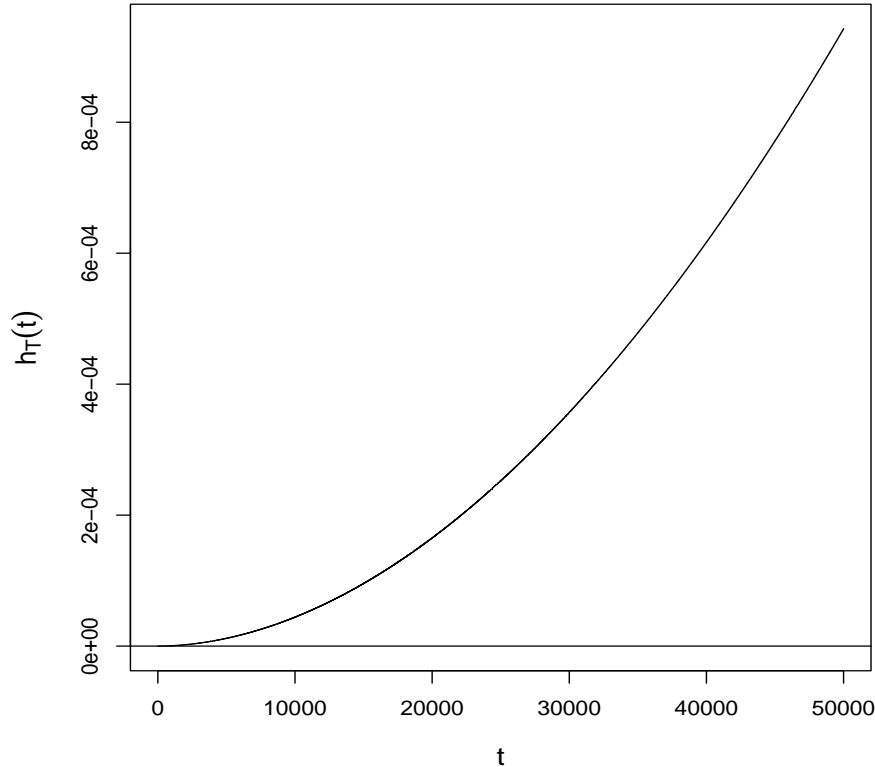


Figure 5.7: (Estimated) hazard function of T in Example 5.2.

Hazard:

$$h_T(t) = \frac{2.9}{19125.6} \left(\frac{t}{19125.6} \right)^{1.9}.$$

This function is shown in Figure 5.7 (above). Because the hazard function is increasing, this means the population of shock absorbers gets weaker over time.

5.4 Quantile-quantile plots

Importance: In a reliability analysis, we will typically assume a lifetime random variable T has a specific distribution, like the Weibull distribution. How do we know if this assumption is reasonable?

- Because we are making an assumption about the distribution of all units (individuals) in the population, we never get to know for sure if the distribution we have chosen is correct.
- We *can* assess if the distribution we have chosen is “reasonable” based on the observed data in the sample.

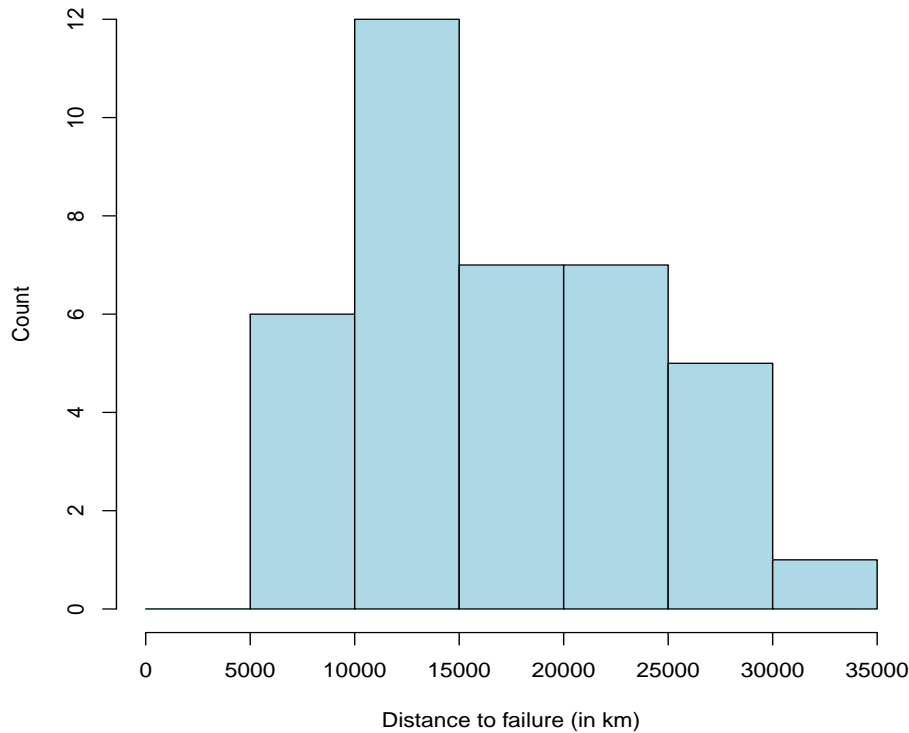


Figure 5.8: Histogram of the shock absorber data in Example 5.2.

- This is part of the “model diagnostics” phase of any data analysis. This means we are assessing (or diagnosing) the plausibility of the assumptions made as part of the analysis.

Remark: The first thing I do in any data analysis is look at the data graphically. A histogram of the $n = 38$ shock absorber distances is shown in Figure 5.8 (above). With such a small sample, it’s hard to make good prognostications about “what’s going on” in the population of all shock absorbers. However, the shape we see in the histogram does align with the right-skewed shape we know is characteristic of a Weibull pdf. This is reassuring but by no means determinative. There are many types of skewed right distributions.

Terminology: A **quantile-quantile plot (qq plot)** is a graphical display that can help assess how well a distribution fits a data set. Here is how the plot is constructed:

- On the vertical axis, we plot the observed data ordered from low to high.
- On the horizontal axis, we plot the same number of (ordered) quantiles from the distribution assumed for the observed data.

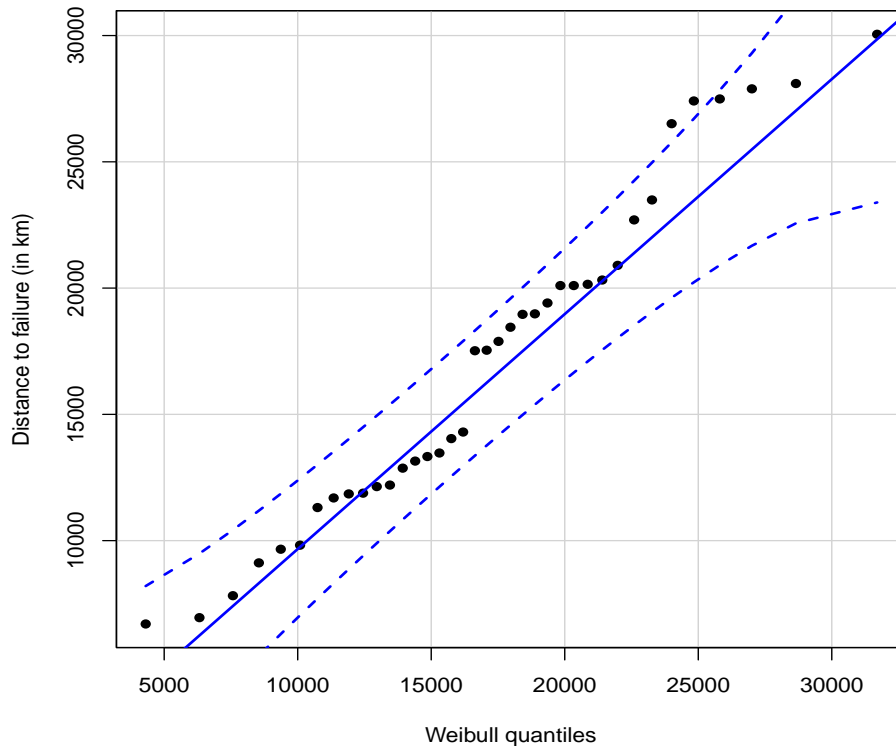


Figure 5.9: Quantile-quantile plot of the shock absorber data in Example 5.2. This is a plot of the observed data (vertical axis) versus quantiles of a Weibull distribution with shape $\hat{\beta} = 2.9$ and scale $\hat{\eta} = 19125.6$ (horizontal axis).

Our intuition should suggest the following:

- If the observed data align with the distribution’s quantiles, then the qq plot should look like a **straight line**. This suggests the distribution fits the data well and is therefore a reasonable choice for the population.
- If the observed data do not align with the distribution’s quantiles, then the qq plot should have **curvature** in it. This suggests the distribution may not be a good choice for the population.

Assessment: The qq plot in Figure 5.9 (above) looks linear for the most part. Even though the agreement isn’t perfect, the Weibull distribution appears to be reasonable for the shock absorber data in Example 5.2.

Important: When you interpret qq plots, you are looking for **general agreement**. The observed data will never line up perfectly with the distribution’s quantiles due to natural variability—even when the distribution is correct! In other words, don’t be “too picky” when interpreting these plots, especially with small sample sizes (like $n = 38$).

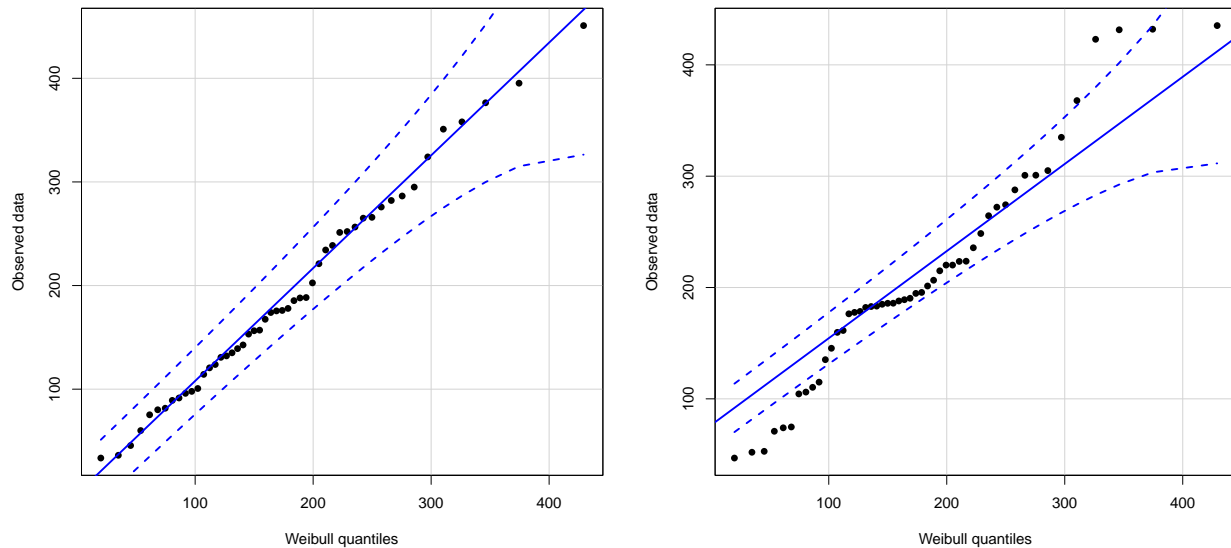


Figure 5.10: Two qq plots with sample size $n = 50$. The Weibull(2, 200) distribution is assumed for the population.

- If there is disagreement, it will usually happen in the tails of the distribution (left or right). With time-to-event data, disagreement in the right tail is more common.
- The bands in Figure 5.9 are shown to give the viewer a reference for how much variability about the line “is allowed.” However, even if a couple of points fall outside the bands, especially if these points are in the tail, this should not cause dramatic concern.

Exercise: Figure 5.10 (above) shows two qq plots, each with sample size $n = 50$. The population distribution is assumed to be Weibull with $\beta = 2$ and $\eta = 200$.

Q: I simulated two data sets. I then constructed qq plots (above) for each data set under the assumption the Weibull(2, 200) population distribution is correct. Which plot do you think used data simulated from the correct distribution?

A: This was a trick question! Both qq plots show data sets simulated from the correct distribution.

- In reality, I simulated about 20 data sets from the correct Weibull(2, 200) distribution and constructed qq plots for each one.
- I then selected the qq plot that looked “the best” (left) and the one that looked “the worst” (right).
- The lesson here is that qq plots, while helpful in model assessment, should not be meticulously overanalyzed. This is especially true with small sample sizes.

6 Bridge to Statistical Inference

6.1 Populations and samples (Parameters and statistics)

Goal: We now shift our focus to **statistical inference**. This deals with making statements about a population of individuals based on information that is available in a sample taken from the population.

- In most situations, it is not possible to observe all individuals in a population (e.g., all power supply units, all shaft bearings, all shock absorbers, etc.). The population is too large, and it would be too time consuming to measure every individual in it.
- If the observed sample is representative of the population, then what we see in the sample should approximate “what’s going on” in the population.
- In this class, we will assume the sample of individuals is a **random sample**. Mathematically, this means all observations are independent and follow the same probability distribution.
- Selecting a random sample is our best hope of obtaining individuals that are representative of the entire population.

Notation: We will denote a random sample of observations by using random variable notation:

$$X_1, X_2, \dots, X_n.$$

That is, X_1 is the value of X for the first individual in the sample, X_2 is the value of X for the second individual in the sample, and so on. The **sample size** tells us how many individuals are in the sample and is denoted by n . Lower case notation x_1, x_2, \dots, x_n is used when citing numerical values. We will typically call these **data**.

Example 6.1. Aluminum-lithium alloys are primarily used in the aerospace industry due to their high strength-to-weight ratio and stiffness. The data below are the compressive strengths (in psi) of $n = 80$ specimens of a new alloy undergoing evaluation as a possible material for aircraft structural elements.

105	221	183	186	121	181	180	143	97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110	163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123	134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169	199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135	196	201	200	176	150	170	118	149

Population: all alloy specimens (of this type) produced using the current process

Sample: the 80 specimens

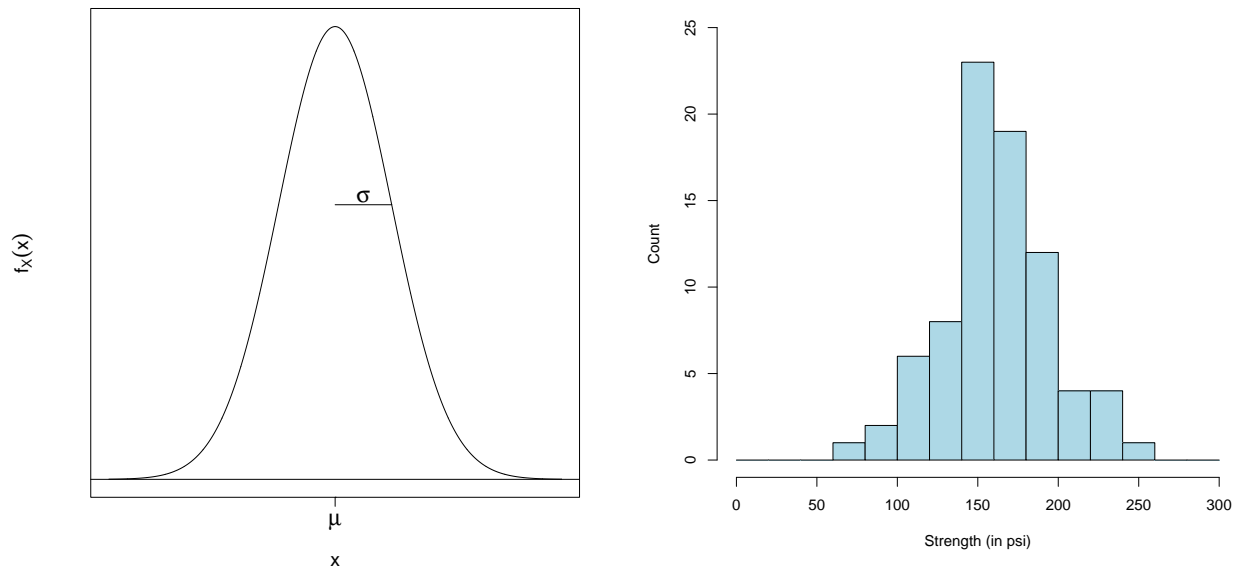


Figure 6.1: Left: $\mathcal{N}(\mu, \sigma^2)$ pdf for X in Example 6.1. This serves as a model for the population of all alloy specimens. Right: Histogram of the sample of $n = 80$ alloy specimens.

Engineers assume the random variable

$$X = \text{strength of alloy specimen (in psi)}$$

is normally distributed with mean μ and variance σ^2 . That is, $X \sim \mathcal{N}(\mu, \sigma^2)$. This is called the **population distribution**. We use the term “population distribution” to mean the distribution of X for all individuals (alloy specimens) in the population. The data, shown above in Figure 6.1 (right), can be conceptualized as 80 independent observations from this population distribution.

Remark: This example illustrates a common situation encountered in practice. The engineers are willing to assume the population distribution is $\mathcal{N}(\mu, \sigma^2)$, but the **population parameters** associated with this distribution

$$\begin{aligned}\mu &= \text{population mean} \\ \sigma^2 &= \text{population variance}\end{aligned}$$

are unknown. The statistical inference question then becomes, “How do we **estimate** these parameters with the observed data?”

Remark: This is analogous to our discussion in the last chapter where we assumed a Weibull(β, η) population distribution for a lifetime random variable T ; see Example 5.2. We assumed a Weibull(β, η) distribution for the population of all shock absorbers. We then estimated the population parameters β and η with the $n = 38$ shock absorbers in the sample and used qq plots to assess the Weibull assumption.

Terminology: A **parameter** is a numerical quantity that describes a population (more specifically, all individuals in the population). In most situations, population parameters are unknown. Some common examples are:

$$\begin{aligned}\mu &= \text{population mean} \\ \sigma^2 &= \text{population variance} \\ \sigma &= \text{population standard deviation} \\ p &= \text{population proportion.}\end{aligned}$$

Terminology: Suppose X_1, X_2, \dots, X_n is a random sample from a population with mean μ and variance σ^2 .

- The **sample mean** is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- The **sample variance** is

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

- The **sample standard deviation** is the positive square root of the sample variance; i.e.,

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

Terminology: A **statistic** is a numerical quantity that is calculated from a sample of observations. In statistical inference, (sample) statistics are used to estimate (population) parameters. We say

- the sample mean \bar{X} is an estimate of the population mean μ
- the sample variance S^2 is an estimate of the population variance σ^2
- the sample standard deviation S is an estimate of the population standard deviation σ .

Example 6.1 (continued). We use R to calculate statistics for the alloy strength data:

```
> options(digits=4)
> mean(strength) # sample mean
[1] 162.7
> var(strength) # sample variance
[1] 1141
> sd(strength) # sample standard deviation
[1] 33.77
```

Therefore, we would conclude:

- the population mean μ can be estimated using $\bar{x} = 162.7$ psi
- the population variance σ^2 can be estimated using $s^2 = 1141$ (psi)²
- the population standard deviation σ can be estimated using $s = 33.77$ psi.

Discussion: It is important to understand that when we calculate an estimate of a population parameter, that's all we are doing—we are taking a “guess” at what it is. There is no guarantee we are correct or even close for that matter.

- Different samples will give different statistic values. For example, if engineers in Example 6.1 sampled another $n = 80$ alloy specimens the following day, they would get different strength measurements and, thus, all statistics' values would change.
- Statistics' values will change from sample to sample. On the other hand, population parameters do not change. They continue to describe the entire population regardless of how many times we sample from it.
- One desirable characteristic of a statistic, in general, is that it estimates the population parameter “correctly on average.” This does not mean one sample will estimate the parameter correctly (some samples will underestimate; some samples will overestimate). This means that over the long run, if one took many samples, the statistic would estimate the parameter correctly on average. This is the definition of **unbiasedness**.
- It also makes sense to think about how variable a statistic's value might be from sample to sample. Doing this will help us understand how much variability is associated with the statistics we calculate. In turn, this will help us form **confidence intervals** for parameters we wish to estimate (next chapter).

6.2 Point estimation and sampling distributions

Note: The ideas in this section can be applied to a variety of situations. Therefore, to keep our discussion general, we let θ denote an arbitrary population parameter.

- For example, θ could denote a population mean, a population variance, a population proportion, a Weibull population distribution parameter, etc.
- It could also denote a parameter in a linear regression model (Chapters 10-11) or other statistical model.
- Whatever the quantity θ represents, the salient point is that it is unknown because it describes the entire population. We want to estimate it using a random sample.

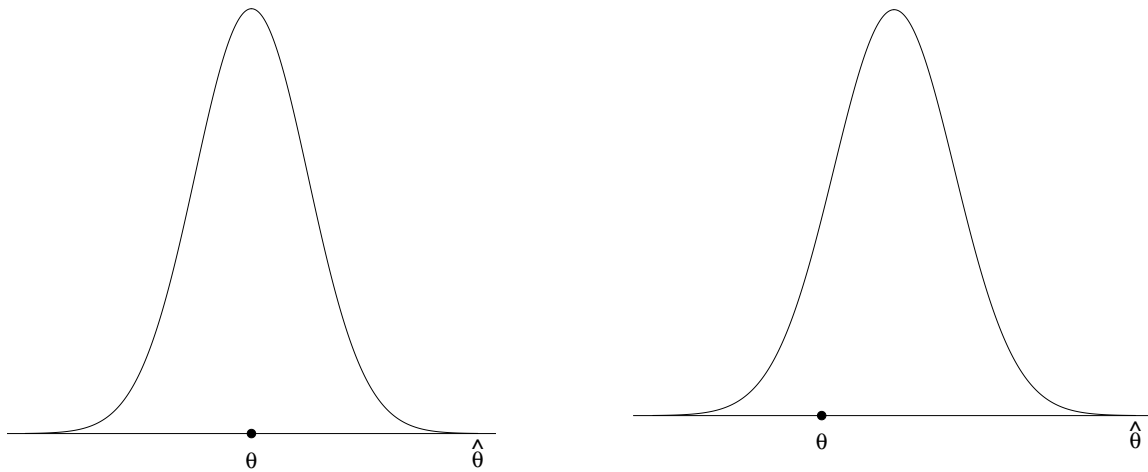


Figure 6.2: Sampling distribution of the point estimator $\hat{\theta}$. Left: $\hat{\theta}$ is an unbiased estimator. Right: $\hat{\theta}$ is a biased estimator.

Terminology: A **point estimator** $\hat{\theta}$ is a statistic that estimates a population parameter θ . Common examples are:

- $\bar{X} \longrightarrow$ a point estimator for the population mean μ
- $S^2 \longrightarrow$ a point estimator for the population variance σ^2
- $S \longrightarrow$ a point estimator for the population standard deviation σ .

Important: Because a point estimator $\hat{\theta}$ is a statistic, its value depends on the sample that is observed, and, as we just discussed, its value will be different for different samples. Therefore, it makes sense to think about the distribution of all possible values of $\hat{\theta}$ that could arise from sampling.

Terminology: The distribution of a point estimator $\hat{\theta}$ is called its **sampling distribution**. This distribution describes how $\hat{\theta}$ would vary in repeated sampling from the same population. We say that $\hat{\theta}$ is an **unbiased estimator** of θ if

$$E(\hat{\theta}) = \theta.$$

In other words, the mean of the sampling distribution of $\hat{\theta}$ is equal to θ . This means that $\hat{\theta}$ will estimate θ “correctly on average.”

Result: Suppose X_1, X_2, \dots, X_n is a random sample from a population with mean μ and variance σ^2 . Mathematics can show

$$\begin{aligned} E(\bar{X}) &= \mu \\ E(S^2) &= \sigma^2. \end{aligned}$$

That is, the sample mean \bar{X} is an unbiased estimator of the population mean μ . The sample variance S^2 is an unbiased estimator of the population variance σ^2 .

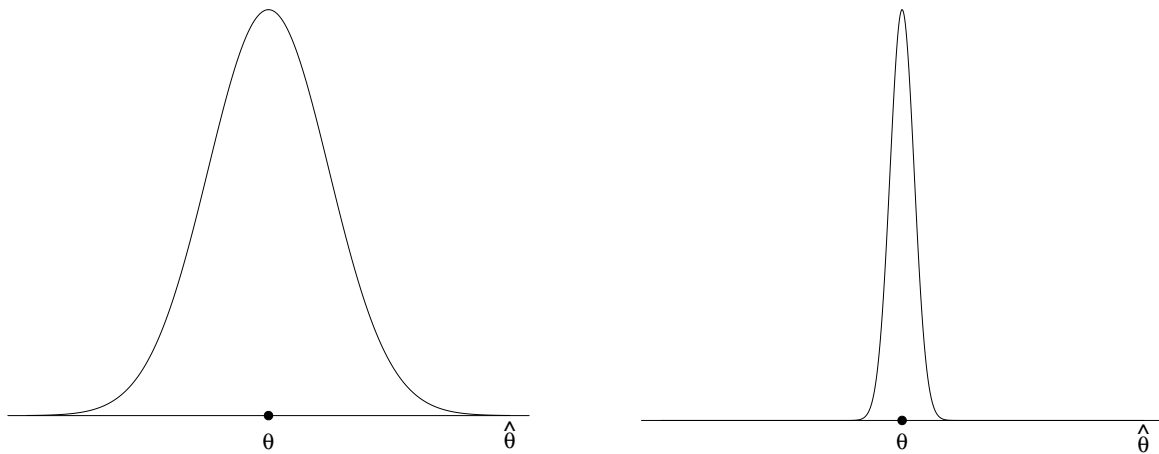


Figure 6.3: Sampling distribution of the point estimator $\hat{\theta}$. The variability associated with $\hat{\theta}$ is smaller for the sampling distribution on the right.

Discussion: Unbiasedness is a desirable property for a point estimator $\hat{\theta}$ to possess. This deals with **accuracy**.

- Unbiased estimators are perfectly accurate. This does not mean $\hat{\theta}$ will estimate θ perfectly for a given sample. Over the long run, $\hat{\theta}$ will estimate θ correctly on average.
- In the light of the last remark, it is important to understand that $\hat{\theta}$ will probably miss θ for a given sample—even when $\hat{\theta}$ is unbiased. By how much will it miss? This is a question about **precision**.
- Figure 6.3 (above) shows two sampling distributions, and $\hat{\theta}$ is an unbiased estimator in both. However, the variability in the sampling distribution on the right is smaller. That is, when $\hat{\theta}$ “misses” θ , it doesn’t miss by as much. The point estimator $\hat{\theta}$ whose sampling distribution is depicted on the right is more precise.
- **Best of both worlds:** We would prefer point estimators $\hat{\theta}$ to be unbiased (perfectly accurate) and have small variance (highly precise). In practice, improving the precision of a point estimator $\hat{\theta}$ can usually be accomplished by increasing the sample size.

Terminology: The **standard error** of a point estimator $\hat{\theta}$ quantifies how variable it is. Specifically, it equals

$$\text{se}(\hat{\theta}) = \sqrt{V(\hat{\theta})}.$$

In other words, the standard error of $\hat{\theta}$ is the standard deviation of its sampling distribution. Therefore,

$$\text{smaller se}(\hat{\theta}) \iff \hat{\theta} \text{ more precise.}$$

Recall: We have seen standard errors before. In Example 5.2, we estimated a Weibull(β, η) distribution for the distance until failure (in km) for a population of shock absorbers. Here was the R output:

```
> options(digits=3)
> fitdist(distance.to.failure,distr="weibull",method="mle")

Fitting of the distribution ' weibull ' by maximum likelihood
Parameters:
      estimate Std. Error
shape      2.9      0.367
scale 19125.6  1140.512
```

In our new estimation language, we would say

- $\hat{\beta} = 2.9$ is a point estimate for β , the population shape parameter
- $\hat{\eta} = 19125.6$ is a point estimate for η , the population scale parameter
- The R output above also displays the standard errors of both point estimates. These describe how variable the point estimates are.
- Point estimates and standard errors will play an important role in writing confidence intervals. We start this discussion in the next chapter.

6.3 Sampling distribution of \bar{X}

Importance: Averages (sample means) are the most widely used statistics, so it is important to understand how they vary in repeated sampling. For example,

- average yield of a chemical production process
- average time to part failure
- average precipitation level
- average number of defects per piece of raw material.

Result 1: Suppose X_1, X_2, \dots, X_n is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. The sample mean \bar{X} has the following sampling distribution:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

- Our first observation here is

normal population distribution \implies sampling distribution of \bar{X} is also normal.

- This result also reminds us that

$$E(\bar{X}) = \mu,$$

that is, the sample mean \bar{X} is an unbiased estimator of the population mean μ .

- This result shows the standard error of \bar{X} is

$$\text{se}(\bar{X}) = \sqrt{V(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

This reveals the variability in the possible values of \bar{X} depends on

- the population standard deviation σ (for individuals in the population)
- the sample size n .

Implication: Larger samples will reduce the variability associated with \bar{X} as a point estimator of μ . This will lead to more precise estimates. This is also true when the population distribution is non-normal (Result 2; coming up).

Example 6.2. In cardiology, an ejection fraction (EF) measures your heart's ability to pump oxygen-rich blood out to your body. This is measured as a percentage and quantifies the amount of blood pumped out of the left ventricle when your heart contracts. Suppose for a population of healthy male subjects, the ejection fraction X is normally distributed with mean $\mu = 56$ and standard deviation $\sigma = 8$.

Q: What is the population distribution?

A: $X \sim \mathcal{N}(56, 64)$. This is the distribution of EF for all male subjects in the population.

Q: A random sample of $n = 16$ males is selected from the population and the EF is measured on each subject producing X_1, X_2, \dots, X_{16} . What is the sampling distribution of \bar{X} , the sample mean EF?

A: Use Result 1:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \implies \bar{X} \sim \mathcal{N}\left(56, \frac{64}{16}\right) \implies \bar{X} \sim \mathcal{N}(56, 4).$$

The sample mean \bar{X} is normally distributed with mean 56 and variance 4.

Q: In the last part, what is the standard error of \bar{X} ?

A: The standard error of \bar{X} is the standard deviation of its sampling distribution, here,

$$\text{se}(\bar{X}) = \sqrt{4} = 2.$$

Note that this is also

$$\text{se}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{16}} = 2$$

using the formula above.

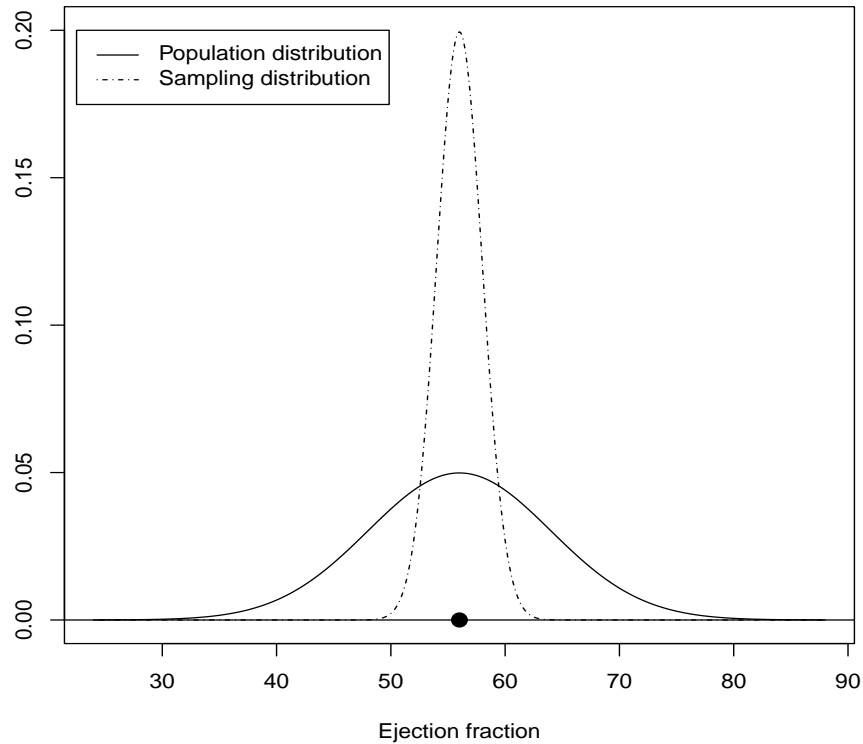


Figure 6.4: $\mathcal{N}(56, 64)$ pdf for X in Example 6.2. This serves as a model for the population of all healthy male subjects. The sampling distribution of \bar{X} with $n = 16$ is also shown.

Q: Calculate $P(X > 60)$ and $P(\bar{X} > 60)$ and explain what these mean.

A: We want

$$P(X > 60) = P\left(\frac{X - 56}{8} > \frac{60 - 56}{8}\right) = P(Z > 0.5) \approx 0.309.$$

This means about 30.9% of the population of all male subjects will have an EF larger than 60. Also,

$$P(\bar{X} > 60) = P\left(\frac{\bar{X} - 56}{8/\sqrt{16}} > \frac{60 - 56}{8/\sqrt{16}}\right) = P(Z > 2) \approx 0.0228.$$

If we observed a random sample of $n = 16$ male subjects from this population, the probability the sample mean ejection fraction \bar{X} would be larger than 60 is about 0.0228.

```
> options(digits=3)
> 1-pnorm(0.5,0,1)
[1] 0.309
> 1-pnorm(2,0,1)
[1] 0.0228
```

Recall: Result 1 informed us that when X_1, X_2, \dots, X_n is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, the sample mean

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Q: What if the population distribution is something else other than a normal distribution? What is the sampling distribution of \bar{X} in this case?

A: We answer this question now, stating one of the most fascinating results in statistics.

Result 2: Suppose X_1, X_2, \dots, X_n is a random sample from a population distribution with mean μ and variance σ^2 . When the sample size n is large, the sample mean

$$\bar{X} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right).$$

The symbol \mathcal{AN} is read “approximately normal.” This result is called the **Central Limit Theorem (CLT)**.

Remark: The CLT implies the sample mean \bar{X} will behave like a normal random variable (approximately) regardless of what the population distribution looks like. The population distribution could be skewed, bimodal, discrete, binary, or whatever. The only mathematical requirement is that the population variance $\sigma^2 < \infty$, which holds for nearly all probability distributions.

Example 6.3. In a textile production process, the number of defects (per square meter) in a certain fabric is assumed to follow a Poisson distribution with mean $\lambda = 1.5$. A quality control plan involves sampling 50 square meter pieces per day and inspecting them for defects.

Q: What is the population distribution?

A: Define

$$X = \text{number of defects per square meter of fabric.}$$

The population distribution is $\text{Poisson}(\lambda = 1.5)$. This is the distribution of the number of defects X for each square meter piece of fabric in the population.

Q: What is the sampling distribution of \bar{X} , the average number of defects for the 50 pieces observed on a given day?

A: The population mean is $\mu = 1.5$ and the population variance is $\sigma^2 = 1.5$. Recall that in the Poisson distribution, the mean and variance are equal. Now, use Result 2 (CLT):

$$\bar{X} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right) \implies \bar{X} \sim \mathcal{AN}\left(1.5, \frac{1.5}{50}\right) \implies \bar{X} \sim \mathcal{AN}(1.5, 0.03).$$

Q: Assuming the population distribution is correct, how likely would it be to observe a sample mean \bar{X} larger than 2 as part of the daily quality control plan?

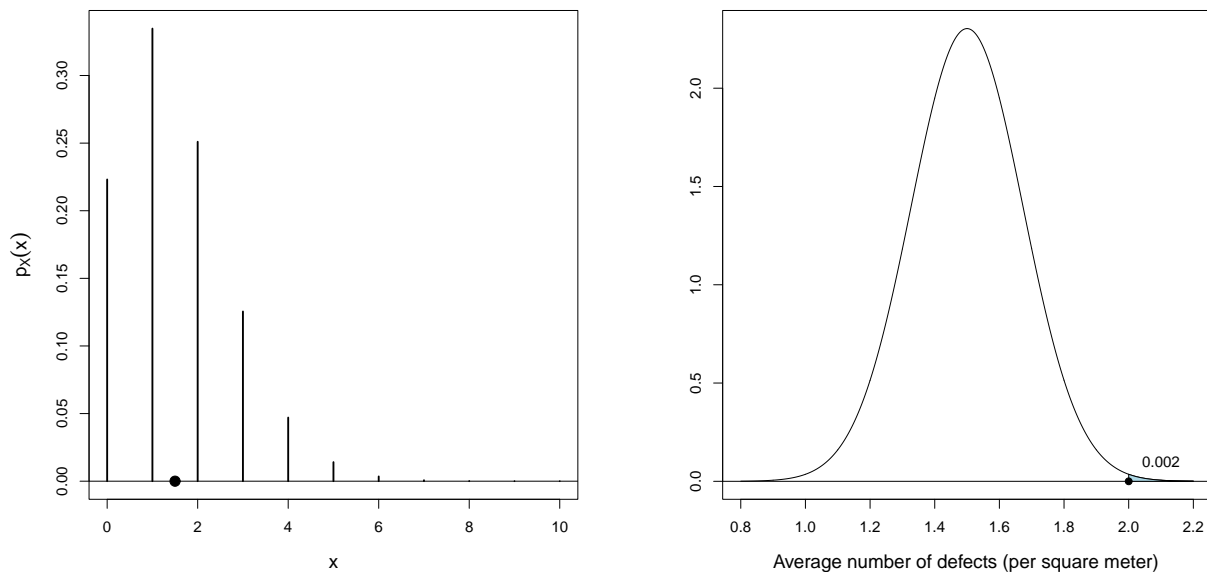


Figure 6.5: Left: Population distribution of $X \sim \text{Poisson}(\lambda = 1.5)$. The population mean $\mu = 1.5$ is shown using a solid circle. Right: Approximate sampling distribution of \bar{X} , the sample mean of $n = 50$ observations from the population. The probability $P(\bar{X} > 2) \approx 0.002$ is shown shaded.

A: We can calculate

$$P(\bar{X} > 2) = P\left(\frac{\bar{X} - 1.5}{\sqrt{1.5/50}} > \frac{2 - 1.5}{\sqrt{1.5/50}}\right) \approx P(Z > 2.89) \approx 0.002.$$

```
> options(digits=1)
> 1-pnorm(2.89,0,1)
[1] 0.002
```

Question for thought: Because $P(\bar{X} > 2) \approx 0.002$ is so small, what might be true if we actually observed a sample mean \bar{X} larger than 2 on any given day? This would certainly not be expected if the population distribution was correct.

Back to the CLT: Because the CLT only approximates the sampling distribution of \bar{X} , that is,

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

it is natural to wonder how good the approximation actually is. This depends primarily on two factors:

1. the sample size n . The larger the sample size, the better the approximation.
2. the amount of skewness in the population distribution. The closer the population distribution is to being **symmetric**, the better the approximation.

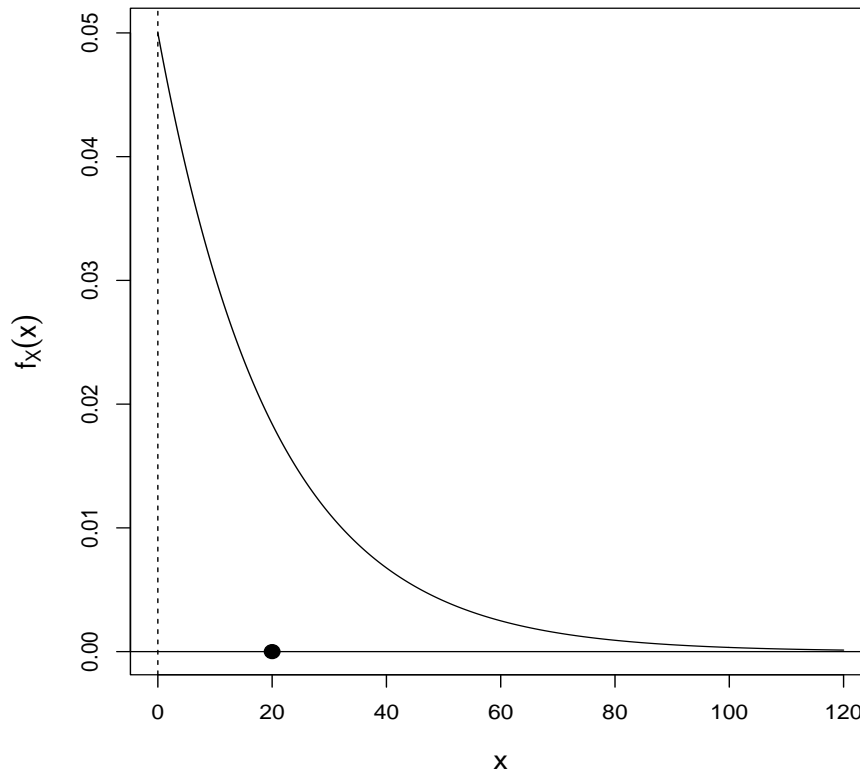


Figure 6.6: Population distribution of $X \sim \text{exponential}(\lambda = 0.05)$ in Example 6.4. The population mean $\mu = 20$ is shown using a solid circle.

Remark: There is no “one sample size n that fits all” to ensure the CLT will offer a good approximation (although some textbooks claim $n \geq 30$ is the magic threshold). For population distributions which are symmetric or approximately symmetric, the sample size n doesn’t have to be that large. For severely skewed distributions or distributions with other nonstandard shapes, the sample size might have to be larger.

Misinterpretation: Some students (and not-so-smart researchers) will interpret the CLT as “with a large enough sample, the data should look approximately normal.” This is not correct. When you are sampling from a population, the population distribution remains fixed—it doesn’t change. The fact that you have a larger sample simply means that you have more observations from the population. The CLT is a statement about the sampling distribution of the sample mean \bar{X} ; not the shape of the population from which you are sampling. The population doesn’t “become more normal” when you have larger samples.

Example 6.4. A clinical trial is being planned to test the effectiveness of semaglutide for weight loss in pre-diabetic patients. The time X (in days) to enroll a patient from this population into the trial follows an exponential distribution with $\lambda = 0.05$ so that the

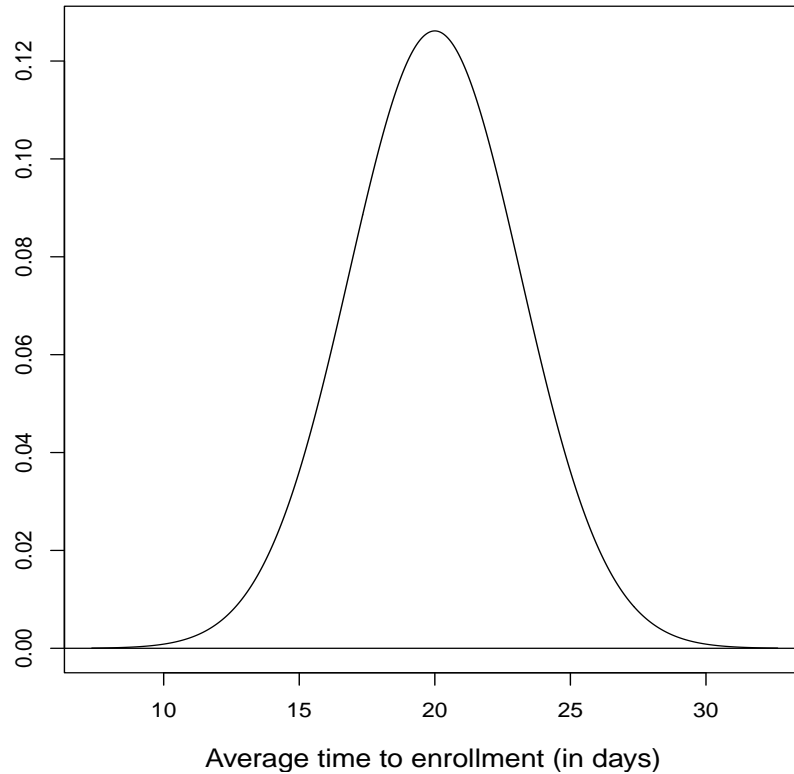


Figure 6.7: Approximate sampling distribution of \bar{X} in Example 6.4.

population mean is

$$\mu = \frac{1}{\lambda} = \frac{1}{0.05} = 20 \text{ days.}$$

The population variance is

$$\sigma^2 = \frac{1}{\lambda^2} = \frac{1}{(0.05)^2} = 400 \text{ (days)}^2.$$

The goal is recruit 40 patients.

Q: What is the population distribution?

A: $X \sim \text{exponential}(\lambda = 0.05)$. This is the distribution of the time to enrollment for individual patients in this population.

Q: What is the sampling distribution of \bar{X} , the sample mean time to enrollment for the 40 patients?

A: Use Result 2 (CLT):

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \implies \bar{X} \sim \mathcal{N}\left(20, \frac{400}{40}\right) \implies \bar{X} \sim \mathcal{N}(20, 10).$$

This sampling distribution is shown in Figure 6.7 (above).

6.4 The t distribution

Recall: Suppose X_1, X_2, \dots, X_n is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. Result 1 informed us that

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Therefore, if we standardize \bar{X} , that is, subtract its mean and divide through by its standard deviation (standard error), we obtain

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

The random variable Z follows a **standard normal distribution**.

New result: If we replace the population standard deviation σ with the sample standard deviation S in the quantity above, we get a new distribution:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

a **t distribution** with $n-1$ degrees of freedom. This result will be used in the next chapter when we discuss confidence intervals for a population mean μ . More generally, the t distribution is widely used in statistical inference.

Facts: The t distribution has the following characteristics:

- Its pdf is continuous and symmetric about 0 (just like the standard normal pdf); see Figure 6.8 (next page).
- It is indexed by a value ν called the **degrees of freedom**. In practice, ν is usually an integer that depends on the sample size.
- When compared to the standard normal pdf, the t pdf is less peaked and has more probability (area) in the tails.
- As $\nu \rightarrow \infty$, the t pdf approaches the standard normal pdf. For $\nu \geq 30$ or so, it is very hard to distinguish the t pdf from the standard normal pdf with the naked eye.

MEAN/VARIANCE: If $T \sim t(\nu)$, then

$$\begin{aligned} E(T) &= 0 \\ V(T) &= \frac{\nu}{\nu-2}. \end{aligned}$$

The mean is $E(T) = 0$, provided that $\nu > 1$. The variance formula above is only applicable when $\nu > 2$.

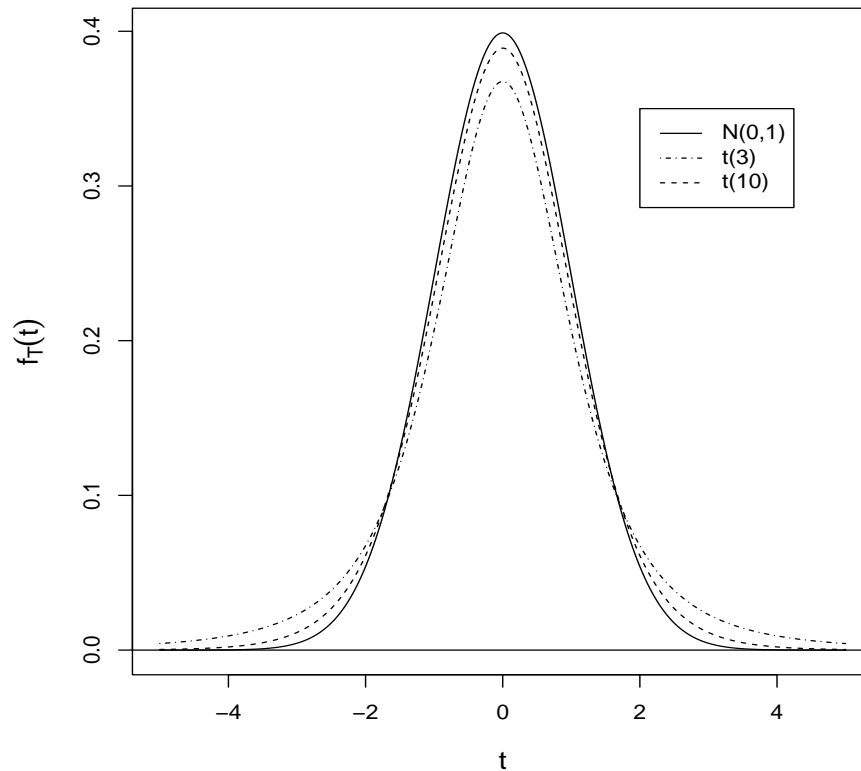


Figure 6.8: t pdfs with $\nu = 3$ and $\nu = 10$ degrees of freedom. The $\mathcal{N}(0, 1)$ pdf is shown as a reference.

Remark: There is a t pdf formula, but it is complicated and unnecessary for our purposes. R will compute probabilities and quantiles from any t distribution.

t R CODE: Suppose $T \sim t(\nu)$.

$$\frac{F_T(t) = P(T \leq t)}{\text{pt}(t, \nu)} \quad \frac{\phi_p}{\text{qt}(p, \nu)}$$

For example, here are the 95th percentiles of each distribution in Figure 6.8 above:

```
> options(digits=3)
> qt(0.95,3) # 95th percentile of t(3)
[1] 2.35
> qt(0.95,10) # 95th percentile of t(10)
[1] 1.81
> qnorm(0.95,0,1) # 95th percentile of N(0,1)
[1] 1.64
```

6.5 Normal quantile-quantile plots

Recall: In the last chapter, we used **quantile-quantile (qq) plots** to assess whether the $\text{Weibull}(\beta, \eta)$ population distribution was appropriate for a lifetime data set. We can make similar plots to assess whether a normal population distribution is appropriate.

Importance: We have just learned that when X_1, X_2, \dots, X_n is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, the quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

In other words, for the t distribution to result, we are assuming the population distribution (from which the data arise) is normal. Normality is a common assumption with many statistical inference techniques we will learn going forward. We need to have a way to assess whether this assumption is reasonable.

Recall: We can use qq plots to assess the normality assumption for the observed data in a sample. Recall how this plot is constructed:

- On the vertical axis, we plot the observed data ordered from low to high.
- On the horizontal axis, we plot the same number of (ordered) quantiles from the population distribution assumed for the observed data (here, a normal distribution).

Linearity in the qq plot supports the normal population assumption. A strong departure from linearity (e.g., extreme curvature) refutes it. Remember, we are looking for **general agreement** when we examine these plots.

Example 6.1 (continued). We observed a sample of 80 alloy specimens and measured the compressive strength (in psi) of each specimen:

105	221	183	186	121	181	180	143	97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110	163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123	134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169	199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135	196	201	200	176	150	170	118	149

Q: Is it reasonable to assume these data arise from a normal population distribution?

A: We can construct a qq plot for the data to answer this question; see Figure 6.9 (next page).

- The plot reveals some minor departures in both tails (lower and upper), but nothing that is too extreme.
- The plot is mostly supportive of the normality assumption for the population of alloy specimens; at least, there isn't strong evidence to refute normality here.

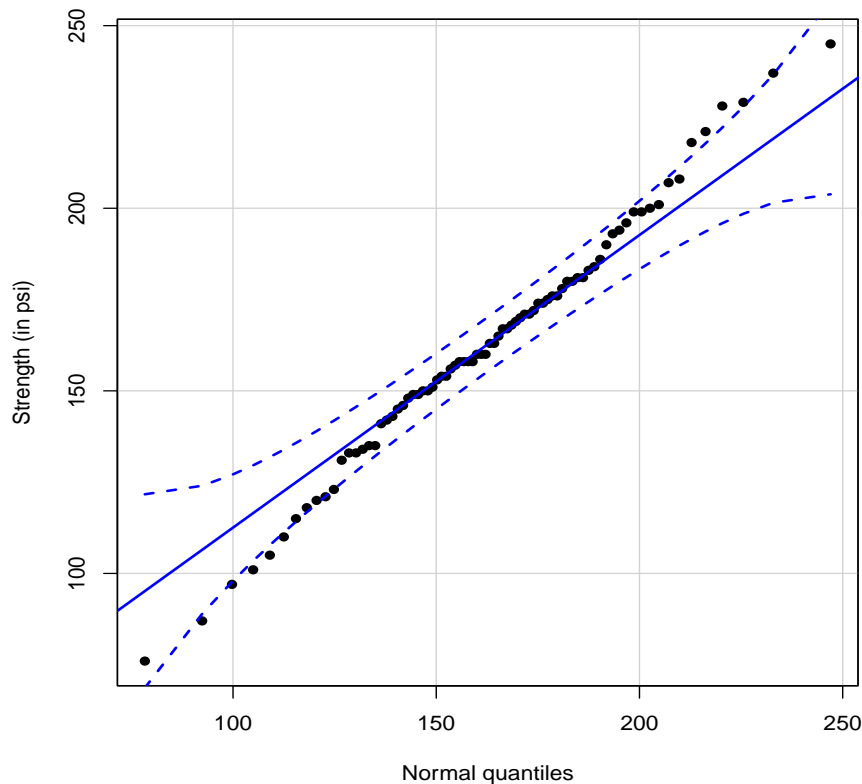


Figure 6.9: Normal quantile-quantile plot of the alloy strength data in Example 6.1.

Example 6.5. Arsenic (As) is a chemical element found naturally in ground water. Excessive levels may result from contamination caused by hazardous waste or by industries that make or use arsenic. Environmental engineers sampled $n = 102$ water wells in Texas and measured the arsenic concentration X (in parts per billion, ppb) for each well. The observed data are shown below:

17.6	10.4	13.5	4.0	19.9	16.0	12.0	12.2	11.4	12.7	3.0	10.3	21.4	19.4	9.0
6.5	10.1	8.7	9.7	6.4	9.7	63.0	15.5	10.7	18.2	7.5	6.1	6.7	6.9	0.8
73.5	12.0	28.0	12.6	9.4	6.2	15.3	7.3	10.7	15.9	5.8	1.0	8.6	1.3	13.7
2.8	2.4	1.4	2.9	13.1	15.3	9.2	11.7	4.5	1.0	1.2	0.8	1.0	2.4	4.4
2.2	2.9	3.6	2.5	1.8	5.9	2.8	1.7	4.6	5.4	3.0	3.1	1.3	2.6	1.4
2.3	1.5	4.0	1.8	2.6	3.4	1.4	10.7	18.2	7.7	6.5	12.2	10.1	6.4	10.7
6.1	0.8	12.0	28.1	9.4	6.2	7.3	9.7	62.1	15.5	6.4	9.5			

Q: Is it reasonable to assume these data arise from a normal population distribution?

A: Figure 6.10 (next page) shows the histogram and the normal qq plot for these data.

- The histogram shows a strong skewed right shape with multiple outliers, which we know doesn't align with the normal distribution.

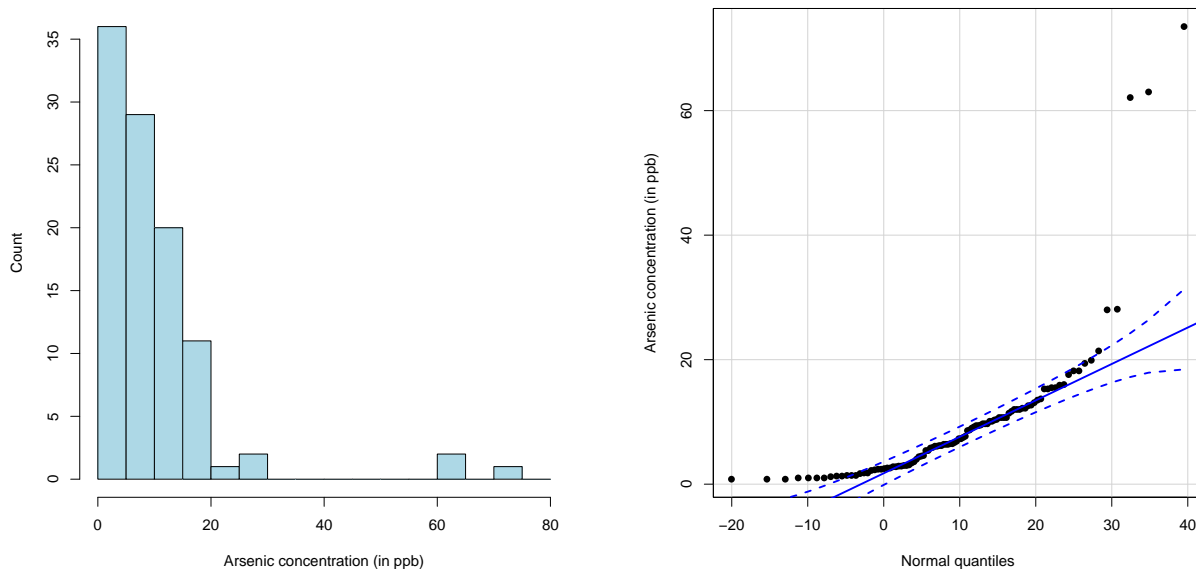


Figure 6.10: Left: Histogram of the $n = 102$ arsenic concentrations in Example 6.5. Right: Normal quantile-quantile plot.

- The appearance of the histogram appears to align more with an exponential or gamma population distribution.
- Not surprisingly, we see **strong disagreement** in the observed data and the normal quantiles in the qq plot. Normality is not a good assumption for these data.

Robustness: We know when X_1, X_2, \dots, X_n is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, the quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

Q: What if X_1, X_2, \dots, X_n is a random sample from some other population distribution (e.g., Poisson, exponential, etc.)? Does the sampling distribution for T above still hold?

A: No, it doesn't, but it may be approximately correct even when the population distribution is not normal. The approximation is best when

- the sample size n is larger
- the population distribution is more symmetric.

Terminology: In statistical inference, we say that a result or method is **robust** when the accuracy of the result (or success of the method) does not depend critically on the underlying assumptions. The t sampling distribution result above is robust to the normality assumption for the population distribution. This means the normality assumption isn't that critical, especially when the two conditions above are met.

7 One-Sample Inference

Preview: In this chapter, we will discuss one-sample inference for three population parameters:

- a population **mean** μ (Section 7.1)
- a population **variance** σ^2 (Section 7.2)
- a population **proportion** p (Section 7.3).

Remember these quantities describe an entire population, so they are unknown. Our goal is to use sample information to estimate them. This is what statistical inference is all about. To “infer” means to “draw a conclusion about something based on evidence.”

Example 7.1. A general contractor buys standard-sized bricks from a local brick supplier. The manufactured specifications call for each brick to weigh 4.5 lbs, but there has been recent concern on the contractor’s part the supplier is selling bricks that do not conform to this specified weight on average. The contractor asks for a sample of bricks to be selected and each brick weighed, producing the data below.

4.54	4.64	4.58	4.78	4.58	4.62	4.55	4.63	4.51	4.49	4.50	4.51
4.63	4.47	4.36	4.61	4.53	4.45	4.26	4.40	4.48	4.63	4.47	4.46
4.57	4.41	4.50	4.62	4.50	4.61	4.49	4.79	4.39	4.70	4.39	4.45

There are $n = 36$ bricks in the sample, but this is a small collection when compared to the tens of thousands of bricks produced each day by the supplier. We can think of the population here as all bricks manufactured by the supplier using the current production process.

- One statistical inference question is, “What are the plausible values of the population mean brick weight μ that are consistent with the data in the sample?” Is 4.5 lbs included within this range of plausible values? If not, then it would appear there is something wrong with the brick supply in terms of the average weight.
- Another equally important question deals with variability. The target specification is 4.5 lbs, but we notice from the data above there is variation in the sample observations (some weights are below 4.5 lbs; some are above). Is there “too much” weight variation in the population of all bricks? A range of plausible values of σ^2 , the population variance, would be helpful in assessing whether there is excessive weight variability in the brick supply.

Statistical formulation: How do we answer these questions? That is what we will do in this chapter. We begin by assuming a random sample of observations X_1, X_2, \dots, X_n is available from a population described by a normal distribution with mean μ and variance σ^2 . Our goal is to use the evidence in the sample to formulate (or “to infer”) a range of plausible values for these population parameters.

7.1 Confidence interval for a population mean μ

Recall: From the last chapter, we learned that if X_1, X_2, \dots, X_n is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

a t distribution with $n-1$ degrees of freedom. This describes the sampling distribution of T , that is, how T will vary probabilistically when sampling from a $\mathcal{N}(\mu, \sigma^2)$ population.

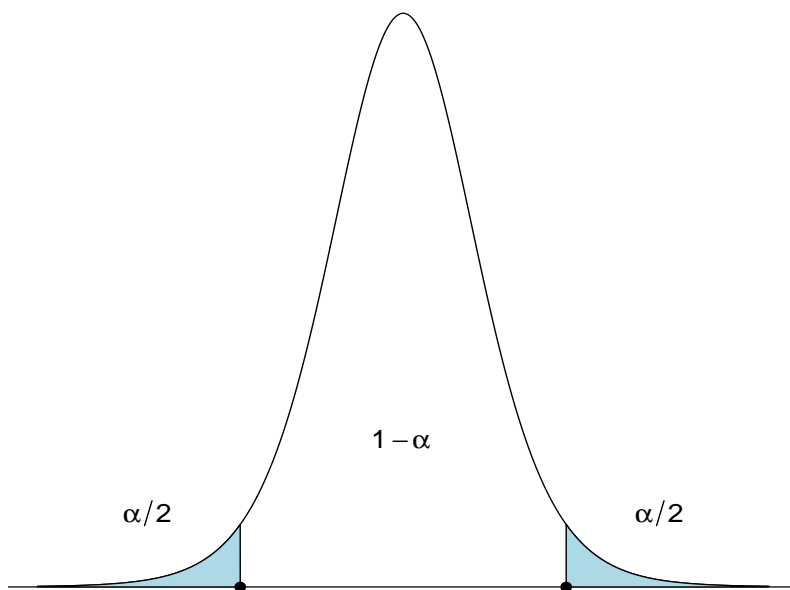


Figure 7.1: t pdf with $n-1$ degrees of freedom. The lower $\alpha/2$ quantile $-t_{n-1, \alpha/2}$ and the upper $\alpha/2$ quantile $t_{n-1, \alpha/2}$ are shown using solid circles.

Notation: We introduce notation that identifies quantiles from a $t(n-1)$ distribution. Define

$$\begin{aligned} t_{n-1, \alpha/2} &= \text{upper } \alpha/2 \text{ quantile from } t(n-1) \text{ pdf} \\ -t_{n-1, \alpha/2} &= \text{lower } \alpha/2 \text{ quantile from } t(n-1) \text{ pdf.} \end{aligned}$$

Because the $t(n-1)$ pdf is symmetric about zero, these two quantiles are equal in absolute value (the upper quantile is positive; the lower quantile is negative); see Figure 7.1. For example, if $n = 10$ and $\alpha = 0.05$, then

$$\begin{aligned} t_{9, 0.025} &\approx 2.26 \\ -t_{9, 0.025} &\approx -2.26. \end{aligned}$$

We can obtain quantiles like these using the `qt` function in R:

```
> options(digits=3)
> qt(0.975,9) # upper 0.025 quantile
[1] 2.26
> qt(0.025,9) # lower 0.025 quantile
[1] -2.26
```

Derivation: For any value of α , $0 < \alpha < 1$, we can write

$$\begin{aligned}
 1 - \alpha &= P\left(-t_{n-1,\alpha/2} < T < t_{n-1,\alpha/2}\right) \quad \leftarrow \text{this comes from Figure 7.1.} \\
 &= P\left(-t_{n-1,\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{n-1,\alpha/2}\right) \\
 &= P\left(-t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} < \bar{X} - \mu < t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}\right) \\
 &= P\left(t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} > \mu - \bar{X} > -t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}\right) \\
 &= P\left(\bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} > \mu > \bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}\right) \\
 &= P\left(\bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}\right).
 \end{aligned}$$

We call

$$\left(\bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}\right)$$

a $100(1 - \alpha)\%$ **confidence interval** for the population mean μ . This is written more succinctly as

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}.$$

Example 7.1 (continued). Calculate a 95% confidence interval for the population mean brick weight μ using the sample data in Example 7.1:

4.54	4.64	4.58	4.78	4.58	4.62	4.55	4.63	4.51	4.49	4.50	4.51
4.63	4.47	4.36	4.61	4.53	4.45	4.26	4.40	4.48	4.63	4.47	4.46
4.57	4.41	4.50	4.62	4.50	4.61	4.49	4.79	4.39	4.70	4.39	4.45

Sample statistics: We use R to first find the sample mean \bar{x} and the sample standard deviation s for these data:

```
> options(digits=3)
> mean(bricks) # sample mean
[1] 4.53
> sd(bricks) # sample standard deviation
[1] 0.113
```

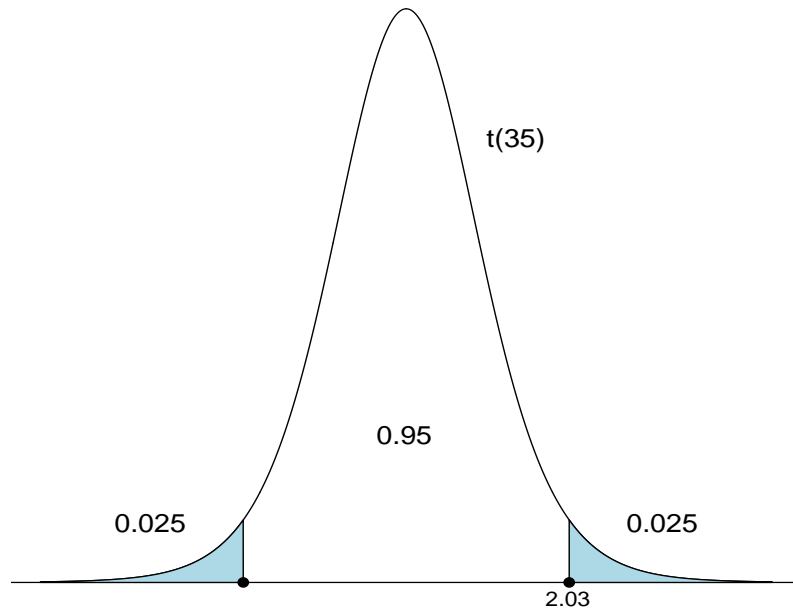
Therefore, we have

$$\begin{aligned}\bar{x} &\approx 4.53 \text{ lbs} \\ s &\approx 0.113 \text{ lbs.}\end{aligned}$$

Quantiles: With a sample of size $n = 36$, we use the $t(35)$ distribution. Note that

$$95\% \text{ confidence} \implies \alpha = 0.05 \implies \alpha/2 = 0.025 \implies t_{35,0.025} \approx 2.03.$$

```
> qt(0.975,35)
[1] 2.03
```



A 95% confidence interval for the population mean μ is

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \implies 4.53 \pm 2.03 \left(\frac{0.113}{\sqrt{36}} \right) \implies (4.49, 4.57).$$

Interpretation: We are 95% confident the population mean brick weight μ is between 4.49 and 4.57 lbs.

Implementation in R: We can calculate confidence intervals for a population mean μ using the `t.test` function in R:

```
> options(digits=3)
> t.test(bricks,conf.level=0.95)$conf.int
[1] 4.49 4.57
```

Using this function avoids the piecemeal approach outlined above, although it is helpful to see all the parts that go into the calculation of the interval.

Discussion: We now make several remarks about the confidence interval

$$\bar{X} \pm t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}$$

for a population mean μ . Many of these remarks apply to other types of confidence intervals we will learn later.

- First, note the form of the interval:

$$\underbrace{\bar{X}}_{\text{point estimate}} \pm \underbrace{t_{n-1, \alpha/2}}_{\text{quantile}} \times \underbrace{\frac{S}{\sqrt{n}}}_{\text{standard error}}.$$

Other confidence intervals we will learn have this same form.

- Here is how we interpret the interval:

“We are $100(1 - \alpha)\%$ confident the population mean μ is in this interval.”

- Unfortunately, the word “confident” does not mean “probability.”
 - The word “confidence” means if we sampled from the population over and over again, each time calculating a $100(1 - \alpha)\%$ confidence interval for the population mean μ , then $100(1 - \alpha)\%$ of the intervals we calculated would contain μ .
- In other words, “confidence” refers to “long term behavior” of many intervals; not probability for the interval we calculated. For example, in Example 7.1, it would **not** be correct to write

$$P(4.49 < \mu < 4.57) = 0.95$$

or say “the population mean brick weight μ is between 4.49 and 4.57 lbs with probability 0.95.” Remember, the population mean μ is a fixed number—it isn’t random. It does not make sense to assign probabilities to events that are not random.

- Unfortunately, there is no way to tell if the confidence interval we calculated contains μ or not. Remember, μ is a population-level parameter so it is unknown. The only way we could determine if the interval contains μ would be to observe every individual in the population. Of course, in this unrealistic scenario, we could determine μ exactly so there would be no need to estimate it with a confidence interval.
- Standard confidence levels are
 - 90% ($\alpha = 0.10$)
 - 95% ($\alpha = 0.05$) ← the most common
 - 99% ($\alpha = 0.01$)

The larger the confidence level, the larger the “long term percentage” of intervals that will contain the population mean μ . Larger confidence levels will produce wider intervals to guarantee this.

- We have all heard the aphorism, “There is no free lunch.”
 - If you use a larger confidence level, then what you get is a larger percentage of confidence intervals that will contain μ . The price you pay is that your interval will be wider and, therefore, less precise.
 - Smaller confidence levels will produce a narrower interval which is more precise. The price you pay is that a smaller percentage of confidence intervals will actually contain μ .
 - Therefore, when writing confidence intervals, there is always a tradeoff between “confidence” and “precision.”

Assumptions: Statistical inference procedures (like confidence intervals) are derived from certain assumptions. It is important to know what these assumptions are, how critical they are, and how to check them. The confidence interval

$$\bar{X} \pm t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}$$

for a population mean μ is created under the following assumptions:

1. X_1, X_2, \dots, X_n is a random sample from the population
2. The population distribution is $\mathcal{N}(\mu, \sigma^2)$.

It is under these assumptions that

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

and this was the starting point to derive the interval above.

- The random sampling assumption is critical. We must assume all observations are coming from the same population distribution and that the observations are mutually independent. This ensures our sample will be representative of the population.
- The normality assumption for the population is not so critical. Remember, we learned in the last chapter the sampling distribution result for T above is **robust** to the normality assumption for the population.
 - This means we can still use the confidence interval formula above even when the normal population assumption does not hold exactly.
 - We can check the normality assumption by using qq plots with the observed sample. As long as there are no **serious departures** from normality detected in the plot, the confidence interval above will likely operate close to the nominal confidence level.
 - Even if there are serious departures from normality, the effect of this is usually small when the sample size n is large. This is a consequence of the CLT.

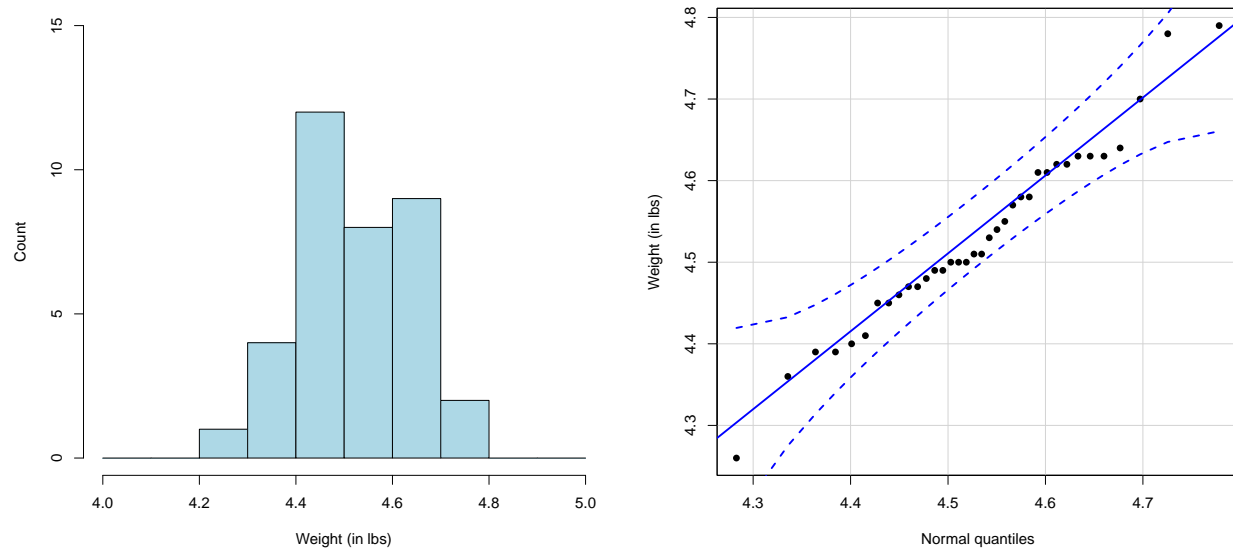


Figure 7.2: Left: Histogram of the $n = 36$ brick weights in Example 7.1. Right: Normal quantile-quantile plot.

Brick weight data: The qq plot for the brick weight data in Figure 7.2 (above, right) does not reveal any serious departures from normality. We can feel comfortable reporting $(4.49, 4.57)$ as a 95% confidence interval for μ , the population mean brick weight.

7.2 Confidence interval for a population variance σ^2

Remark: In some situations, we aren't concerned with the mean of the population but the variance instead. This is especially true in manufacturing settings. Too much variation can lead to inconsistent product quality and reduced customer satisfaction. On the other hand, small levels of variation are associated with predictability. Ensuring predictable results in production and construction is essential for meeting specifications and maintaining a positive reputation with customers.

New result: If X_1, X_2, \dots, X_n is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, then

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

a χ^2 **distribution** with $n-1$ degrees of freedom. This describes the sampling distribution of Q , that is, how Q will vary probabilistically when sampling from a $\mathcal{N}(\mu, \sigma^2)$ population.

Importance: The χ^2 distribution (and this sampling distribution result above) will be used to develop a confidence interval for the population variance σ^2 .

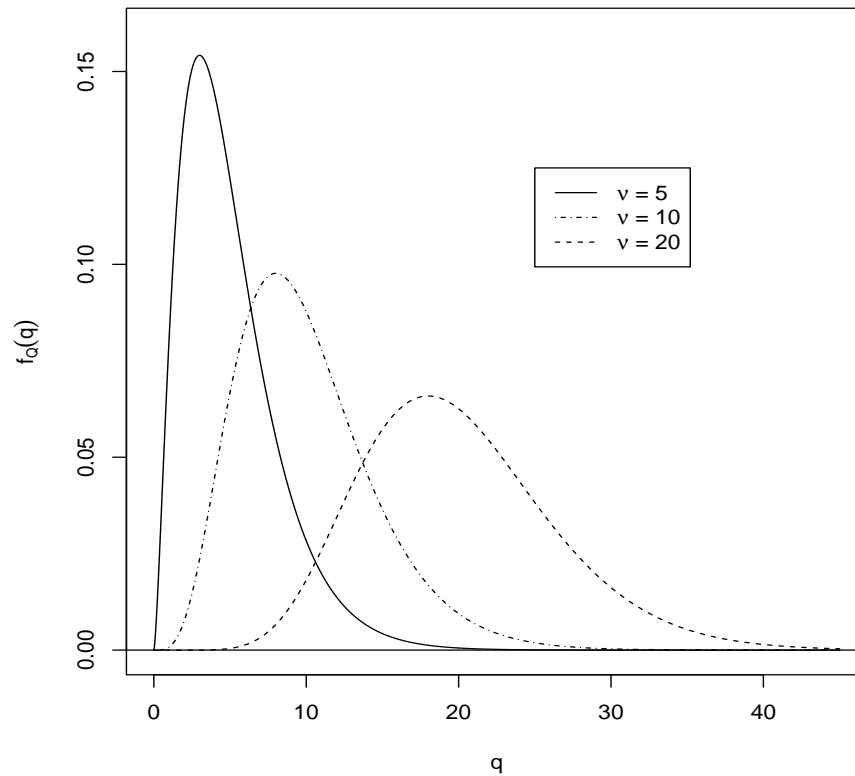


Figure 7.3: χ^2 pdfs with different degrees of freedom.

Facts: The χ^2 distribution has the following characteristics:

- Its pdf is continuous, skewed right, and its support is positive values only (no negative values); see Figure 7.3 above.
- It is indexed by a value ν called the **degrees of freedom**. In practice, ν is usually an integer that depends on the sample size.
- The χ^2 pdf formula is unnecessary for our purposes. R will compute probabilities and quantiles from any χ^2 distribution.

χ^2 **R CODE:** Suppose $Q \sim \chi^2(\nu)$.

$F_Q(q) = P(Q \leq q)$	ϕ_p
<code>pchisq(q, ν)</code>	<code>qchisq(p, ν)</code>

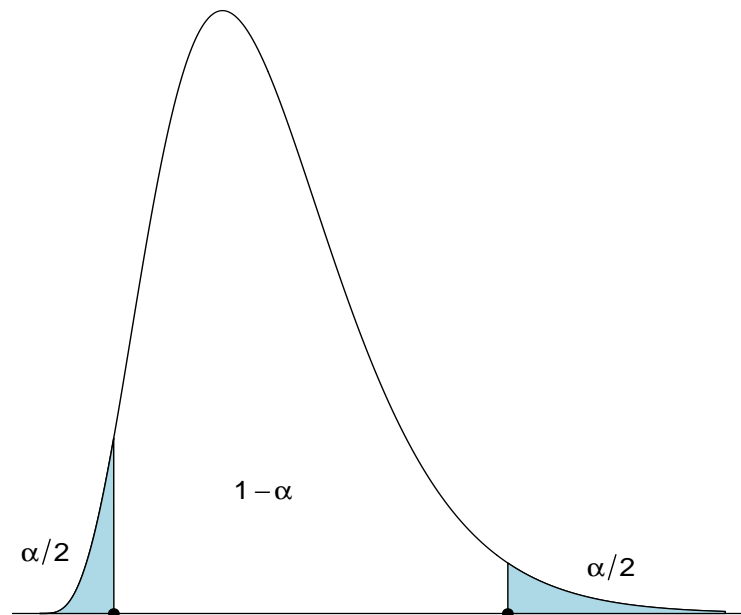


Figure 7.4: χ^2 pdf with $n - 1$ degrees of freedom. The lower $\alpha/2$ quantile $\chi^2_{n-1,\alpha/2}$ and the upper $\alpha/2$ quantile $\chi^2_{n-1,1-\alpha/2}$ are shown using solid circles.

Notation: We introduce notation that identifies quantiles from a $\chi^2(n - 1)$ distribution. Define

$$\begin{aligned}\chi^2_{n-1,1-\alpha/2} &= \text{upper } \alpha/2 \text{ quantile from } \chi^2(n - 1) \text{ pdf} \\ \chi^2_{n-1,\alpha/2} &= \text{lower } \alpha/2 \text{ quantile from } \chi^2(n - 1) \text{ pdf.}\end{aligned}$$

For example, if $n = 10$ and $\alpha = 0.05$, then

$$\begin{aligned}\chi^2_{9,0.975} &\approx 19.02 \\ \chi^2_{9,0.025} &\approx 2.7.\end{aligned}$$

We can obtain quantiles like these using the `qchisq` function in R:

```
> options(digits=4)
> qchisq(0.975,9) # upper 0.025 quantile
[1] 19.02
> qchisq(0.025,9) # lower 0.025 quantile
[1] 2.7
```

Derivation: For any value of α , $0 < \alpha < 1$, we can write

$$\begin{aligned}
 1 - \alpha &= P\left(\chi_{n-1, \alpha/2}^2 < Q < \chi_{n-1, 1-\alpha/2}^2\right) \quad \leftarrow \text{this comes from Figure 7.4.} \\
 &= P\left(\chi_{n-1, \alpha/2}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{n-1, 1-\alpha/2}^2\right) \\
 &= P\left(\frac{1}{\chi_{n-1, \alpha/2}^2} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi_{n-1, 1-\alpha/2}^2}\right) \\
 &= P\left(\frac{(n-1)S^2}{\chi_{n-1, \alpha/2}^2} > \sigma^2 > \frac{(n-1)S^2}{\chi_{n-1, 1-\alpha/2}^2}\right) \\
 &= P\left(\frac{(n-1)S^2}{\chi_{n-1, 1-\alpha/2}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{n-1, \alpha/2}^2}\right).
 \end{aligned}$$

This shows

$$\left(\frac{(n-1)S^2}{\chi_{n-1, 1-\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1, \alpha/2}^2} \right)$$

is a $100(1 - \alpha)\%$ **confidence interval** for the population variance σ^2 . We interpret the interval in the same way:

“We are $100(1 - \alpha)\%$ confident the population variance σ^2 is in this interval.”

Note: A $100(1 - \alpha)\%$ confidence interval for the **population standard deviation** σ arises from simply taking the square root of the endpoints of the σ^2 interval.

- That is,

$$\left(\sqrt{\frac{(n-1)S^2}{\chi_{n-1, 1-\alpha/2}^2}}, \sqrt{\frac{(n-1)S^2}{\chi_{n-1, \alpha/2}^2}} \right)$$

is a $100(1 - \alpha)\%$ confidence interval for the population standard deviation σ .

- This is usually preferred over the σ^2 interval, because standard deviation measures variability in terms of the original units (e.g., dollars, inches, days, etc.).
- Recall the variance σ^2 is measured in squared units (e.g., dollars², in², days², etc.) and is harder to interpret.

Example 7.2. A furniture company sells items designed for the customer to assemble him or herself. One item uses screws which are supposed to have a mean diameter of 1.200 cm. Periodically, quality control is performed to assess whether there is excessive variation in various screw dimensions. Specifications mandate the population standard deviation σ of the diameters should not exceed 0.005 cm. Otherwise, there is excessive variation in the production of this critical part which could lead to difficulty in construction and customer dissatisfaction.

Below are diameter measurements for a random sample $n = 40$ screws:

1.194	1.177	1.204	1.195	1.209	1.208	1.210	1.206	1.187	1.187
1.219	1.198	1.196	1.194	1.194	1.208	1.207	1.201	1.214	1.203
1.198	1.195	1.207	1.185	1.189	1.191	1.204	1.199	1.196	1.212
1.198	1.188	1.203	1.199	1.211	1.215	1.202	1.206	1.212	1.189

Q: Find a 99% confidence interval for the population standard deviation σ .

A: We will first find a 99% confidence interval for the population variance σ^2 and then take the square root of each interval endpoint.

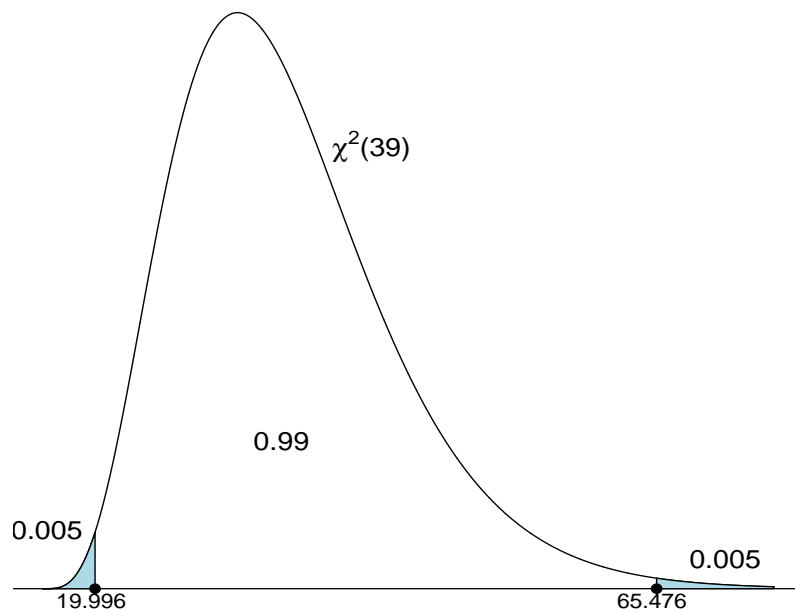
Sample statistics: We use R to first find the sample variance s^2 for these data:

```
> options(digits=5)
> var(screws) # sample variance
[1] 8.9372e-05
```

Therefore, we have

$$s^2 \approx 8.9372 \times 10^{-5} \text{ or } 0.000089372 \text{ cm}^2.$$

I'm retaining many digits due to the small numbers here (this will mitigate the impact of rounding error later).



Quantiles: With a sample of size $n = 40$, we use the $\chi^2(39)$ distribution. Note that

$$99\% \text{ confidence} \implies \alpha = 0.01 \implies \alpha/2 = 0.005 \implies \begin{cases} \chi^2_{39,0.005} \approx 19.996 \\ \chi^2_{39,0.995} \approx 65.476 \end{cases}$$

```
> qchisq(0.005,39)
[1] 19.996
> qchisq(0.995,39)
[1] 65.476
```

Therefore, a 99% confidence interval for the population variance σ^2 is

$$\left(\frac{(n-1)s^2}{\chi_{n-1,1-\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1,\alpha/2}^2} \right) \Rightarrow \left(\frac{39(0.000089372)}{65.476}, \frac{39(0.000089372)}{19.996} \right) \\ \Rightarrow (0.00005323, 0.00017431).$$

Interpretation: We are 99% confident the population variance σ^2 of the screw diameters is between 0.00005323 and 0.00017431 cm².

A 99% confidence interval for the population standard deviation σ is

$$(\sqrt{0.00005323}, \sqrt{0.00017431}) \Rightarrow (0.007, 0.013).$$

Interpretation: We are 99% confident the population standard deviation σ of the screw diameters is between 0.007 and 0.013 cm.

Discussion: This analysis suggests there is too much variation in the process producing the screws. The confidence interval consists entirely of values which are larger than 0.005 cm. It would be advised to revisit the production process and see if we can find assignable causes of variability—those causes which are inflating the variation in the diameters beyond the acceptable upper limit of 0.005 cm.

Implementation in R: There is no handy internal function in R that calculates a confidence interval for a population variance σ^2 , so I wrote one. The function `var.ci` below asks the user to input the data set (`data`) and specify the confidence level.

```
var.ci = function(data,conf.level=0.99){
  df = length(data)-1
  chi.lower = qchisq((1-conf.level)/2,df)
  chi.upper = qchisq((1+conf.level)/2,df)
  s2 = var(data)
  c(df*s2/chi.upper,df*s2/chi.lower)
}
```

Applying this function to the screw diameter data, we get the output

```
> options(digits=5)
> var.ci(screws,conf.level=0.99) # CI for population variance
[1] 5.3234e-05 1.7431e-04
> options(digits=1)
> sqrt(var.ci(screws,conf.level=0.99)) # CI for population standard deviation
[1] 0.007 0.013
```

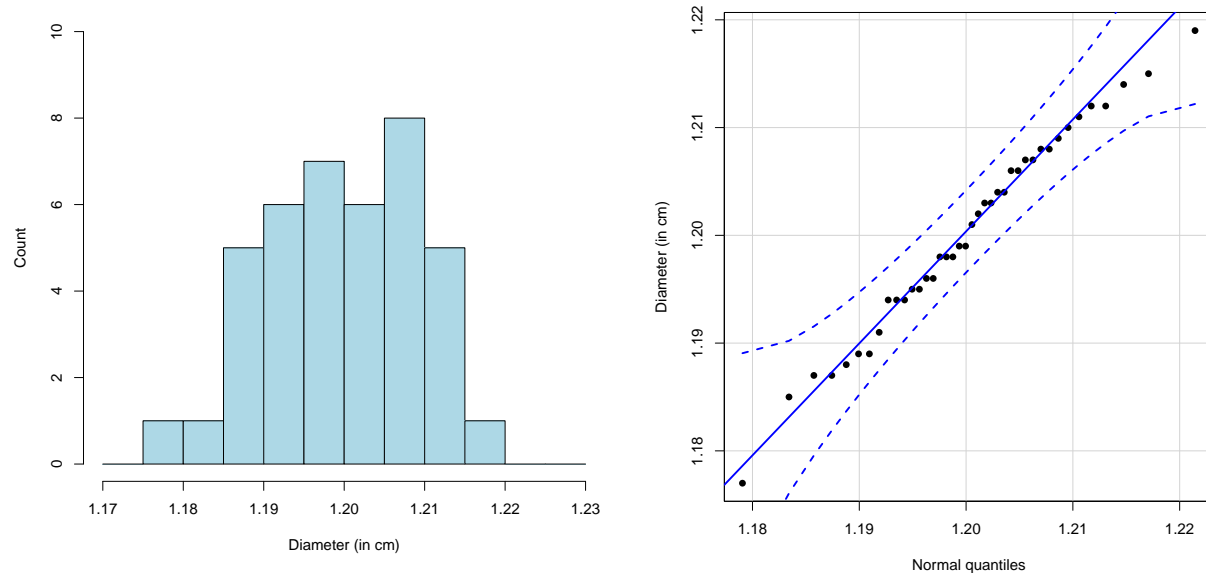



Figure 7.5: Left: Histogram of the $n = 40$ screw diameters in Example 7.2. Right: Normal quantile-quantile plot.

Assumptions: The confidence interval

$$\left(\frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2} \right)$$

for a population variance σ^2 is created under the following assumptions:

1. X_1, X_2, \dots, X_n is a random sample from the population
2. The population distribution is $\mathcal{N}(\mu, \sigma^2)$.

It is under these assumptions that

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

and this was the starting point to derive the interval above.

- The random sampling assumption is critical. We must assume all observations are coming from the same population distribution and that the observations are mutually independent. This ensures our sample will be representative of the population.
- The normality assumption for the population is also critical. The sampling distribution result for Q above is **not** robust to departures from the normality assumption for the population.

- This means the confidence interval for σ^2 (and therefore the one for σ too) is only meaningful when the population is $\mathcal{N}(\mu, \sigma^2)$. Departures from normality could cause the confidence intervals for σ^2 and σ to be misleading.
- This is different than what we learned with a confidence interval for a population mean μ using the t distribution. This interval is robust to the normality assumption.

Screw diameter data: The qq plot for the screw diameter data in Figure 7.5 (last page, right) does not reveal any serious departures from normality. We can feel comfortable reporting (0.007, 0.013) as a 99% confidence interval for σ , the population standard deviation of the diameters.

7.3 Confidence interval for a population proportion p

Scenario: We now switch gears and focus on estimating a population proportion p . This parameter is relevant when we measure a binary characteristic on each individual. Here are some examples:

- p = proportion of defective circuit boards
- p = proportion of power supply units requiring service during a warranty period
- p = proportion of customers who are satisfied
- p = proportion of payments received on time
- p = proportion of patients who respond to treatment.

To start our discussion, we need to recall the **Bernoulli trial** assumptions for each individual in the population:

1. each individual is categorized as a “success” or “failure”
2. the individuals are independent
3. the probability of “success” p is the same for every individual in the population.

In our examples above,

- “success” \longrightarrow circuit board defective
- “success” \longrightarrow PSU requires service during a warranty period
- “success” \longrightarrow customer satisfied
- “success” \longrightarrow payment received on time
- “success” \longrightarrow patient responds to treatment.

The parameter p is the proportion of “successes” in the population. Our goal is to estimate p with a confidence interval.

Recall: A **binomial distribution** arises when we observe a fixed number of Bernoulli trials and record

$X = \text{number of successes (out of } n\text{)}.$

- In a statistical inference context, we envision each individual (e.g., circuit board, PSU, customer, payment, patient, etc.) in a random sample as a “trial,” and we record a “success” or “failure” on each individual in the sample.
- If p is the proportion of “successes” in the population, then $X \sim b(n, p)$.

Point estimation: If X is the number of “successes” in a random sample of size n , then

$$\hat{p} = \frac{X}{n}$$

is the **sample proportion**. This is simply the proportion of successes in the sample. We will use \hat{p} as a point estimator for the population proportion p . Mathematics can show these two results:

$$\begin{aligned} E(\hat{p}) &= p \\ V(\hat{p}) &= \frac{p(1-p)}{n}. \end{aligned}$$

The first result says the sample proportion \hat{p} is an **unbiased estimator** of the population proportion p . From the second result, we can find the standard error

$$\text{se}(\hat{p}) = \sqrt{V(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}.$$

Recall the standard error of a point estimator (like \hat{p}) measures how much variation is attached to it.

Important: Knowing the sampling distribution of \hat{p} is critical if we are going to develop a confidence interval for p . We appeal to an approximation (conferred by the CLT) which says

$$\hat{p} \sim \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right),$$

when the sample size n is large. Standardizing \hat{p} , that is, subtracting the mean and dividing through by the standard deviation (standard error), we get

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{AN}(0, 1),$$

an approximate **standard normal distribution**. We will use this result to write a confidence interval for p .

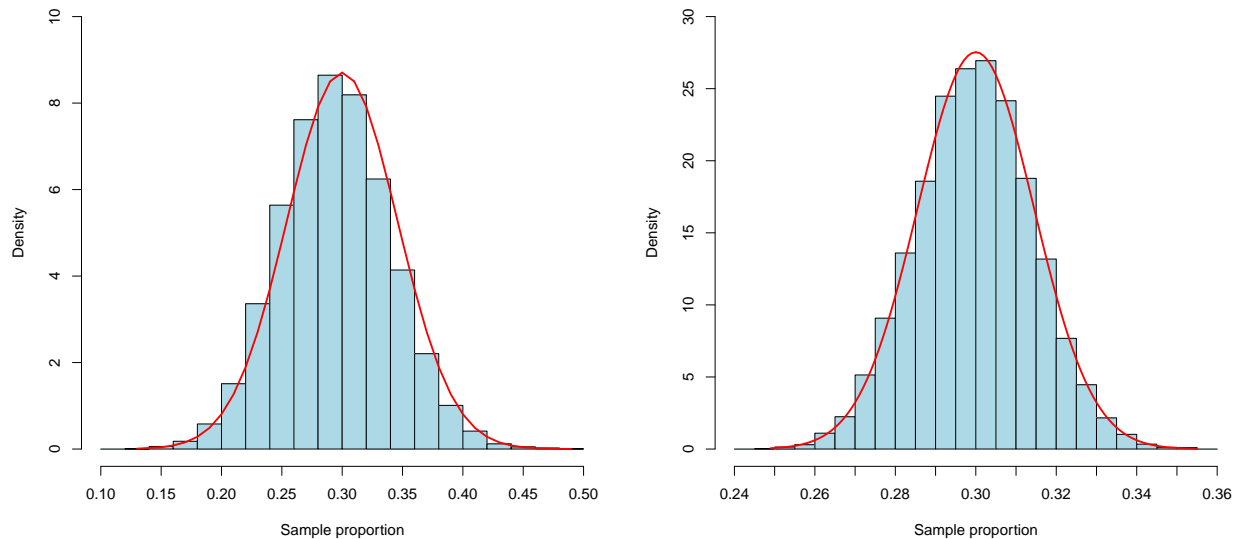


Figure 7.6: Left: Histogram of the 10,000 sample proportions \hat{p} when $n = 100$ and $p = 0.30$. Right: $n = 1000$ and $p = 0.30$. A normal pdf has been superimposed over each histogram. Density histograms have been used so that total histogram areas equal one.

Monte Carlo simulation: Before we develop the confidence interval for p , you should be reasonably convinced the sampling distribution result for \hat{p} on the previous page is correct (at least the approximate normality part).

- The histograms in Figure 7.6 (above) each show 10,000 simulated values of \hat{p} when the population proportion is $p = 0.30$; i.e., 30% of individuals in the population are “successes.”
- The normal approximation to the histogram when $n = 100$ (left) is pretty good. The approximation is outstanding when $n = 1000$ (right). This illustrates how the normal approximation for the sampling distribution of \hat{p} is better for larger sample sizes.
- Notice how the histograms of \hat{p} are centered at $p = 0.30$. This is because \hat{p} is an unbiased estimator.
- Also notice how the variation in the sampling distribution is much smaller when the sample size $n = 1000$ (right). This can be seen also through standard error calculations:

$$\begin{aligned} \text{Left: } \text{se}(\hat{p}) &= \sqrt{\frac{0.30(0.70)}{100}} \approx 0.046 \\ \text{Right: } \text{se}(\hat{p}) &= \sqrt{\frac{0.30(0.70)}{1000}} \approx 0.014. \end{aligned}$$

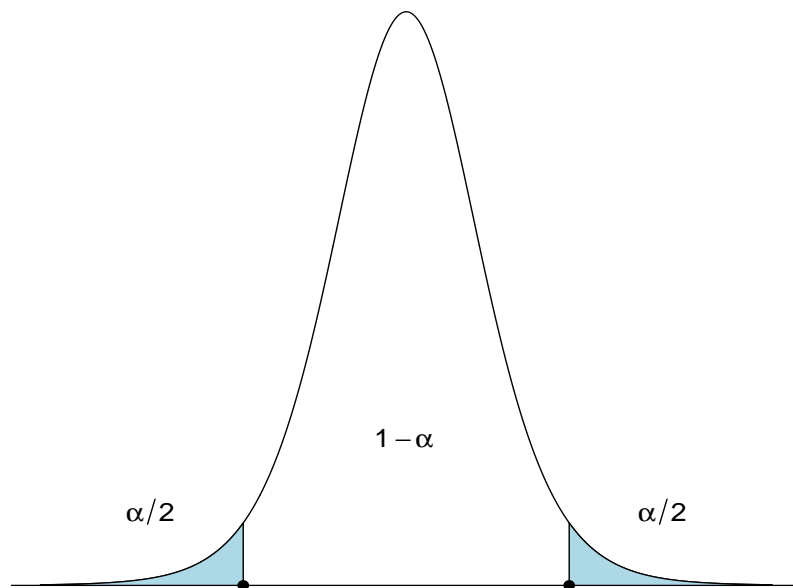


Figure 7.7: $\mathcal{N}(0, 1)$ pdf. The lower $\alpha/2$ quantile $-z_{\alpha/2}$ and the upper $\alpha/2$ quantile $z_{\alpha/2}$ are shown using solid circles.

Notation: We introduce notation that identifies quantiles from a $\mathcal{N}(0, 1)$ distribution. Define

$$\begin{aligned} z_{\alpha/2} &= \text{upper } \alpha/2 \text{ quantile from } \mathcal{N}(0, 1) \text{ pdf} \\ -z_{\alpha/2} &= \text{lower } \alpha/2 \text{ quantile from } \mathcal{N}(0, 1) \text{ pdf.} \end{aligned}$$

Because the $\mathcal{N}(0, 1)$ pdf is symmetric about zero, these two quantiles are equal in absolute value (the upper quantile is positive; the lower quantile is negative); see Figure 7.7. For example, if $\alpha = 0.05$, then

$$\begin{aligned} z_{0.025} &\approx 1.96 \\ -z_{0.025} &\approx -1.96. \end{aligned}$$

We can obtain quantiles like these using the `qnorm` function in R:

```
> options(digits=3)
> qnorm(0.975,0,1)
[1] 1.96
> qnorm(0.025,0,1)
[1] -1.96
```

Derivation: For any value of α , $0 < \alpha < 1$, we can write

$$\begin{aligned}
 1 - \alpha &\approx P(-z_{\alpha/2} < Z < z_{\alpha/2}) \quad \leftarrow \text{this comes from Figure 7.7.} \\
 &\approx P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < z_{\alpha/2}\right) \quad \leftarrow \text{using an estimate of the standard error.} \\
 &= P\left(-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \hat{p} - p < z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \\
 &= P\left(z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} > p - \hat{p} > -z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \\
 &= P\left(\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} > p > \hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \\
 &= P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).
 \end{aligned}$$

We call

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

a $100(1 - \alpha)\%$ **confidence interval** for the population proportion p . This is written more succinctly as

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We interpret the interval in the same way:

“We are $100(1 - \alpha)\%$ confident the population proportion p is in this interval.”

Discussion: Note the familiar form of the interval:

$$\underbrace{\hat{p}}_{\text{point estimate}} \pm \underbrace{z_{\alpha/2}}_{\text{quantile}} \times \underbrace{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{\text{standard error}}.$$

This interval should be used only when the sample size n is “large.” A common **rule of thumb** is to require

$$\begin{aligned}
 n\hat{p} &\geq 5 \\
 n(1 - \hat{p}) &\geq 5.
 \end{aligned}$$

Under these conditions, the CLT should adequately describe the sampling distribution of \hat{p} , thereby making the confidence interval formula above approximately valid.

Example 7.3. One source of water pollution is gasoline leakage from underground storage tanks. In Pennsylvania, a random sample of $n = 74$ gasoline stations is selected from the state and the tanks are inspected; 10 stations are found to have at least one leaking tank.

Q: Calculate a 95% confidence interval for p , the population proportion of gasoline stations in Pennsylvania with at least one leaking tank.

A: In this example, we interpret

- gasoline station = “trial”
- at least one leaking tank at station = “success.”

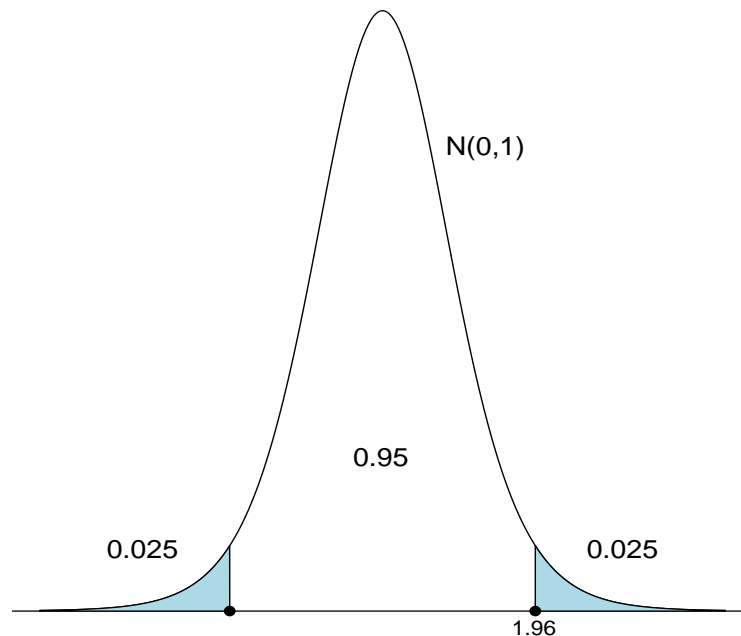
Note: There are about 10,000 gas stations in Pennsylvania. Therefore, p represents the proportion of stations with at least one leaking tank for this entire population. We have a random sample of 74 stations from this population.

Sample statistics: The sample proportion of stations with at least one leaking tank is

$$\hat{p} = \frac{10}{74} \approx 0.135.$$

Quantiles: We use the $\mathcal{N}(0, 1)$ distribution. Note that

$$95\% \text{ confidence} \implies \alpha = 0.05 \implies \alpha/2 = 0.025 \implies z_{0.025} \approx 1.96.$$



```
> options(digits=3)
> qnorm(0.975,0,1)
[1] 1.96
```

A 95% confidence interval for the population proportion p is

$$0.135 \pm 1.96 \sqrt{\frac{0.135(1 - 0.135)}{74}} \implies (0.057, 0.213).$$

Interpretation: We are 95% confident the population proportion p of stations in Pennsylvania with at least one leaking tank is between 0.057 and 0.213.

CLT approximation check: We have

$$\begin{aligned} n\hat{p} &= 74 \left(\frac{10}{74} \right) = 10 \\ n(1 - \hat{p}) &= 74 \left(1 - \frac{10}{74} \right) = 64. \end{aligned}$$

Both of these are larger than 5, so we can feel comfortable using this confidence interval formula.

Implementation in R: There are various R functions and packages which produce confidence intervals for proportions, but they are far more elaborate than what we need. I wrote a simple function `p.ci` that automates the calculations. It requires the user to input the number of successes (`x`), the number of trials (`n`), and the confidence level:

```
p.ci = function(x,n,conf.level=0.95){
  est = x/n
  se = sqrt(est*(1-est)/n)
  z.upper = qnorm((1+conf.level)/2,0,1)
  c(est-z.upper*se,est+z.upper*se)
}
```

Using this function with $x = 10$ and $n = 74$ in Example 7.3, we get

```
> options(digits=2)
> p.ci(10,74,conf.level=0.95)
[1] 0.057 0.213
```

7.4 Sample size determination

Importance: In the planning stages of an experiment or observational study, we need to first determine how many individuals should be sampled from a population. For example,

- We want to write a 90% confidence interval for the population mean time to part failure. How many parts should be sampled?
- We want to write a 95% confidence interval for the population proportion of patients who respond to a treatment. How many patients should we recruit?

The sample size n determines how precise confidence intervals will be. Larger sample sizes will give narrower (more precise) confidence intervals. Of course, sampling from a population always costs money, and larger samples will usually cost more too (think of the Pennsylvania gas station example). We will discuss statistical issues associated with sample size determination. There is a complementary set of practical issues like cost, time spent in sampling, personnel training, and other factors, which are also important.

Sample size for a population mean: Recall that if X_1, X_2, \dots, X_n is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, then

$$\bar{X} \pm t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}$$

is a $100(1 - \alpha)\%$ confidence interval for the population mean μ . This interval is formed by taking a point estimate \bar{X} and then adding/subtracting

$$t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}.$$

We call this quantity the **margin of error** associated with the confidence interval. It consists of two parts:

- the quantile $t_{n-1, \alpha/2}$, which depends on the confidence level $100(1 - \alpha)\%$
- S/\sqrt{n} , which is a point estimate of the population standard error σ/\sqrt{n} . Note the population standard error depends on the population standard deviation σ and the sample size n .

Recall the t distribution is similar to the standard normal distribution $\mathcal{N}(0, 1)$, especially when the degrees of freedom $(n - 1)$ is larger. Therefore, the margin of error above should be approximately equal to

$$B = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Solving this equation for n , we get

$$n = \left(\frac{z_{\alpha/2} \sigma}{B} \right)^2.$$

This is the sample size which will guarantee a prescribed confidence level $100(1 - \alpha)\%$ and a margin of error B .

Note: To use the formula, we have to specify three things:

1. the confidence level $100(1 - \alpha)\%$; this determines the $\mathcal{N}(0, 1)$ quantile $z_{\alpha/2}$.
2. the margin of error B ; note the interval length is twice the margin of error. Therefore, when you specify B , you are specifying how narrow (how precise) you want the interval to be.

3. the population standard deviation σ . This is the hardest part, because σ is a population-level parameter (so it is unknown). Generally, an “estimate” or “guess” is provided here. An upper bound for σ is used if you want to be **conservative**. This will provide a sample size larger than what is actually needed.

Example 7.4. In a biomedical experiment, we would like to estimate the population mean time to death μ for healthy rats (in days) when given a toxic substance. The research protocol requires a 95% confidence interval for μ with a margin of error equal to $B = 2$ days. From past studies, the time to death after administration of the toxin has been modeled by a normal distribution with standard deviation $\sigma = 8$ days.

Q: How many rats should we use for the experiment?

A: With $z_{0.05/2} = z_{0.025} \approx 1.96$, $B = 2$, and $\sigma = 8$, the minimum sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{B} \right)^2 = \left(\frac{1.96 \times 8}{2} \right)^2 \approx 61.5.$$

We would need a random sample of $n = 62$ rats to achieve these goals. This will produce a 95% confidence interval for μ with a margin of error no larger than $B = 2$ days (total interval length no larger than 4 days).

Remark: We could weaken our requirements to (a) a lower 90% confidence level and (b) a margin of error of $B = 3$ days, which is less precise. The minimum sample size is now

$$n = \left(\frac{z_{\alpha/2}\sigma}{B} \right)^2 = \left(\frac{1.65 \times 8}{3} \right)^2 \approx 19.4.$$

We would need to sample only $n = 20$ rats to meet these weaker requirements.

Sample size for a population proportion: A $100(1 - \alpha)\%$ confidence interval for a population proportion p is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

The margin of error associated with the interval

$$z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad \text{is an estimate of} \quad z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}}.$$

A small problem arises, namely, the (population-level) margin of error

$$z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}}$$

depends on p , which is what we are trying to estimate with a confidence interval. The “workaround” is to elicit a “guess” of what p is, say p_0 , and use

$$B = z_{\alpha/2} \sqrt{\frac{p_0(1 - p_0)}{n}}$$

instead. Solving this equation for n , we get

$$n = \left(\frac{z_{\alpha/2}}{B} \right)^2 p_0(1 - p_0).$$

This is the sample size which will guarantee a prescribed confidence level $100(1 - \alpha)\%$ and a margin of error B .

Remark: If there is no sensible guess for p available, use $p_0 = 0.5$. The resulting sample size n will be as large as possible. Put another way, using $p_0 = 0.5$ gives the most **conservative** solution. This is true because

$$n = n(p_0) = \left(\frac{z_{\alpha/2}}{B} \right)^2 p_0(1 - p_0),$$

when viewed as a function of p_0 , is maximized when $p_0 = 0.5$. However, the resulting sample size could be very large, perhaps much larger than is practical to use.

Example 7.5. You have been asked to estimate the proportion of parts in a certain manufacturing process that need to be “scrapped;” e.g., the part is so defective that it can not be used or reworked. If this proportion is larger than 10 percent, it will be deemed by management to be an unacceptable continued operating cost and a substantial process overhaul will be performed. Past experience suggests the scrap rate is about 5 percent, but recent information suggests this rate may be increasing.

You would like to write a 95% confidence interval for p , the population proportion of parts that will be scrapped, with a margin of error equal to $B = 0.02$.

Q: How many parts should you ask to be sampled?

A: For 95% confidence, we use $z_{0.05/2} = z_{0.025} \approx 1.96$. In providing an initial guess p_0 , we have different options; we could use

$$\begin{aligned} p_0 &= 0.05 \text{ (historical scrap rate)} \\ p_0 &= 0.10 \text{ (“critical mass” value)} \\ p_0 &= 0.50 \text{ (most conservative choice).} \end{aligned}$$

For these choices, we have

$$\begin{aligned} n &= \left(\frac{1.96}{0.02} \right)^2 0.05(1 - 0.05) \approx 457 \\ n &= \left(\frac{1.96}{0.02} \right)^2 0.10(1 - 0.10) \approx 865 \\ n &= \left(\frac{1.96}{0.02} \right)^2 0.50(1 - 0.50) \approx 2401. \end{aligned}$$

This shows how the “guess” p_0 can have a substantial impact on the final sample size calculation. Furthermore, it may not be practical to sample 2,401 parts, as would be required by the most conservative approach.