

# STAT243: Intro Stats

S DeRuiter (slightly modified from an original by R Pruim)

Spring 2016

## Contents

<b>1</b>	<b>Probability</b>	<b>3</b>
1.1	Key Definitions and Ideas . . . . .	3
1.2	Calculating Probabilities Empirically . . . . .	5
1.3	Calculating Probabilities Theoretically . . . . .	8
1.4	Conditional Probability . . . . .	12
<b>2</b>	<b>Densities</b>	<b>21</b>
2.1	Density histograms, density plots, density functions . . . . .	21
2.2	Working with Probability Density Functions . . . . .	25
2.3	Some Important Families of Distributions . . . . .	29
2.4	Fitting Distributions to Data . . . . .	39
2.5	Quantile-Quantile Plots . . . . .	47

## 1

## Random Variables and Probability

## 1.1 Key Definitions and Ideas

**random process** A repeatable process that has multiple unpredictable potential outcomes.

Although we sometimes use language that suggests that a *particular result* is random, it is really the *process* that is random, not its results.

**outcome** A potential result of a random process.

**sample space** The set of all possible potential outcomes of a random process.

**event** A subset of the sample space. That is, a set of outcomes (possibly all or none of the outcomes).

Statisticians often use capital letters from the beginning of the alphabet for events.

**trial** One repetition of a random process.

**mutually exclusive** events. Events that cannot happen on the same trial.

**probability** A numerical value between 0 and 1 assigned to an event to indicate how often the event occurs (in the long run).

**random variable** A random variable is a variable whose value is a numerical outcome of a random process.

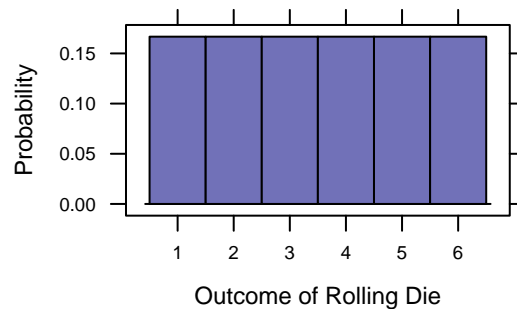
Examples of random variables:

- Roll a die and record the number.
- Roll two dice and record the sum.
- Flip 100 coins and count the number of heads.
- Sample 1000 people and count how many approve of the job the president is doing.
- Sample 100 people and measure how long their right toenail is in mm. (Is this a good example or not? It's a topic for discussion...)

Note: Statisticians usually use capital letters (often from the end of the alphabet) for random variables, like this: Let  $X$  be the number of heads in 10 flips of a fair coin. What is  $P(X = 5)$ ?

**probability distribution** The distribution of a random variable. (Remember that a distribution describes *what values?* and *with what frequency?*)

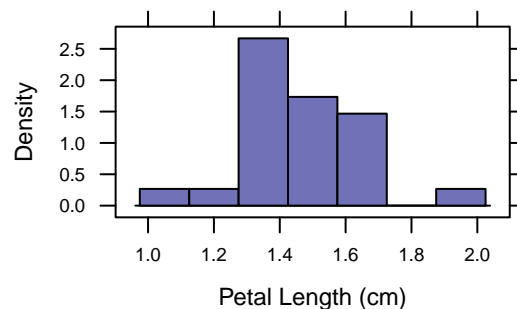
As an example of a probability distribution, we can first consider a *discrete* random variable. Most of the examples of random variables given above are discrete. In other words, the values they can take on come from a set containing a finite number of possible values. For example, if you roll a 6-sided die and record the number that comes up, there are only six possible outcomes, which are equally likely: the integers 1, 2, 3, 4, 5 and 6. For discrete random variables, the probability distribution shows all the possible values on the x-axis, and the likelihood of observing each of those values on the y-axis. Since there are a finite number of possible values that can be observed, these likelihoods are actually the *probabilities* of observing each outcome, and the sum of all the probabilities must be 1. For our example, where we rolled a die and recorded the value:



Things are a bit more complicated for *continuous* random variables (the ones that can take on any numerical value). Here, the sample space (the set of possible distinct values the random variable can take on) is infinite. One consequence of this fact is that the interpretation of the y-axis values of the probability distribution changes. The y-axis will still indicate the relative likelihood of observing any given value of the random variable. However, here the random variable can take on an infinite number of possible values. In this case, we can't interpret the y-axis values as probabilities. The y-axis units are called "Likelihood" or "Density", and they indicate the relative frequency of each outcome.

For a densityplot, which shows a smoothed version of the silhouette of a histogram (take STAT 343 and/or read about Kernel Density Estimation if you want a better explanation), Density is scaled such that the integral over all possible x-values (the area under the curve) is 1. For a histogram, Density is relative frequency, scaled so that the total area of all the boxes added together is 1. We can think of the histograms and density plots we have been creating using continuous variables from R datasets as attempts to use data to approximate the distributions of random variables.

For example, we might consider the growth of flower petals of the iris *Iris setosa* as a random process, and let  $X$  be a random variable that is the length of each iris petal. We could plot a histogram to approximate the distribution of  $X$  using the variable `Petal.Length` from the `iris` data (from the `datasets` package in base R).



## 1.2 Calculating Probabilities Empirically

We would like to calculate the probability of an event  $A$ , denoted  $P(A)$ .

In the next section, we will see how to calculate probabilities based on the Axioms of probability, and logic. But first, we will consider ways to make the calculations empirically – based on observing many repetitions of a random process (in real life or in a computer simulation) and observing how often an event of interest occurs.

Random processes are repeatable, so practically, we can calculate empirical probabilities by simply repeating the process over and over and keeping track of how often the event  $A$  occurs. For example, we could flip a coin 10,000 times and see what fraction are heads.<sup>1</sup>

$$\text{Empirical Probability} = \frac{\text{number of times } A \text{ occurred}}{\text{number of times random process was repeated}}$$

Modern computing provides another way to compute empirical probabilities. If we can simulate our random process on a computer, then we can repeat the process many times very quickly.

**Example 1.2.1.** Q. What is the probability of getting exactly 5 heads if you flip a fair coin 10 times? Using our random variable notation, let  $X$  be the number of heads in 10 flips of a fair coin. We want to know  $P(X = 5)$ .

A. The `rflip()` function simulates flipping a coin as many times as we like.

```
rflip(10)

##
## Flipping 10 coins [ Prob(Heads) = 0.5 ] ...
##
## T H T H T H T H T
##
## Number of Heads: 4 [Proportion Heads: 0.4]
```

The `do()` function allows us to execute an R command ("do" something in R) over and over, as many times as we choose. Here, our `rflip()` command simulates 10 coin-flips. First we'll "do" our command three times and show the results.

Then we'll do it 10,000 times and store the results in a variable called `tosses`, so we can create a table and a plot showing the empirical distribution.

```
do(3) * rflip(10)

##      n heads tails prop
## 1 10      3      7 0.3
## 2 10      7      3 0.7
## 3 10      5      5 0.5

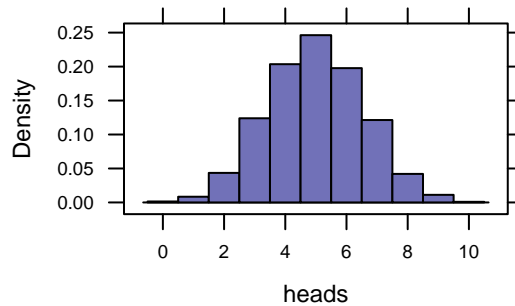
tosses <- do(10000) * rflip(10)
tally(~heads, data = tosses, format = "prop")

##
```

<sup>1</sup>This has actually been done a couple of times in history, including once by mathematician John Kerrich while he was a prisoner of war during World War II.

```
##      0      1      2      3      4      5      6      7      8      9     10
## 0.0014 0.0084 0.0435 0.1239 0.2035 0.2462 0.1977 0.1213 0.0420 0.0112 0.0009
```

```
histogram(~heads, data = tosses, width = 1)
```



Based on this sample, we would estimate that  $P(X = 5) \approx 0.2462$ .

**Example 1.2.2.** Q. Use simulations to estimate the probability of rolling doubles using two fair standard dice.

A. We can simulate rolling a die with the following code:

```
1:6 # the numbers 1 through 6

## [1] 1 2 3 4 5 6

resample(1:6, 10) # ten rolls of a 6-sided die

## [1] 4 5 6 5 2 1 2 6 1 4
```

The first 2 input arguments of `resample()` are `x` (the set of values from which you want to resample) and `size` (the number of items to choose from `x`). You can also think of `size` as the number of *times* to sample from `x`, if you are imagining sampling one item from `x` each time.

If we do this 10,000 times for each of two dice...

```
die1 <- resample(1:6, 10000)
die2 <- resample(1:6, 10000)
# let's check that things look reasonable
head(die1)

## [1] 2 1 5 4 4 1

head(die2)

## [1] 4 2 2 6 2 1
```

Then we can tabulate how often the two numbers matched in one of two ways:

```
tally(~(die1 == die2)) # NOTE the double == here

##
## TRUE FALSE
## 1652 8348

prop(~(die1 == die2)) # NOTE the double == here

## TRUE
## 0.1652
```

So the probability appears to be approximately 0.1652.

**Example 1.2.3.** Q. Use simulation to estimate the probability of rolling a sum of 8 when rolling two fair six-sided dice.

A. We have already generated 10000 random rolls, so let's just reuse them. (Alternatively, we could generate new rolls.)

```
s <- die1 + die2
# R adds element-wise: first entry of die1 + first of die2, second to second, etc.
prop(~(s == 8))

## TRUE
## 0.1357
```

We can estimate the probability of any sum the same way.

```
tally(~s)

##
## 2 3 4 5 6 7 8 9 10 11 12
## 296 555 816 1057 1416 1692 1357 1135 814 588 274

# if we are too lazy to divide by 10000 ourselves:
tally(~s, format = "percent")

##
## 2 3 4 5 6 7 8 9 10 11 12
## 2.96 5.55 8.16 10.57 14.16 16.92 13.57 11.35 8.14 5.88 2.74
```

Here's a slightly fancier version that puts all the information into a data frame. Note the use of the function `data.frame()` to create the data table:

```
rolls <- data.frame(first = die1, second = die2, sum = die1 + die2)
head(rolls)

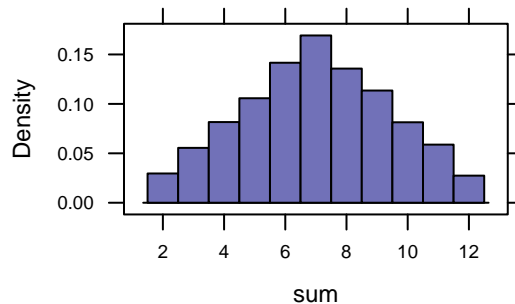
## first second sum
## 1 2 4 6
## 2 1 2 3
```

```
## 3      5      2      7
## 4      4      6     10
## 5      4      2      6
## 6      1      1      2

tally(~sum, data = rolls, format = "proportion")

##
##      2      3      4      5      6      7      8      9     10     11     12
## 0.0296 0.0555 0.0816 0.1057 0.1416 0.1692 0.1357 0.1135 0.0814 0.0588 0.0274

histogram(~sum, data = rolls, width = 1) # setting width is important for integer data
```



## 1.3 Calculating Probabilities Theoretically

The theoretical method combines

1. Some basic facts about probability (the Probability Axioms and Rules),
2. Some assumptions about the particular situation at hand, and
3. Mathematical reasoning (arithmetic, algebra, logic, etc.).

### 1.3.1 The Three Probability Axioms

Let  $S$  be the sample space and let  $A$  and  $B$  be events.

1. Probability is between 0 and 1:  $0 \leq P(A) \leq 1$ .
2. The probability of the sample space is 1:  $P(S) = 1$ .
3. Additivity: If  $A$  and  $B$  are mutually exclusive, then  $P(A \text{ or } B) = P(A) + P(B)$ .

Notation Notes

$P(A \text{ or } B)$  is the probability that either  $A$  or  $B$  (or both) occurs. Often this is written  $P(A \cup B)$ .  $A \cup B$  is usually read “ $A$  union  $B$ ”. The union of two sets is the set that contains all elements of both sets.



$P(A \text{ and } B)$  is the probability that *both*  $A$  and  $B$  occur. This is also written  $P(A \cap B)$ .  $A \cap B$  is usually read “ $A$  intersect  $B$ ”.

Saying that  $A$  and  $B$  are mutually exclusive is the same as saying that there are no outcomes in  $A \cap B$ , i.e., that  $A \cap B = \emptyset$ .

### 1.3.2 Other Probability Rules

These rules all follow from the axioms (although we will not necessarily prove them all here).

#### The Addition Rule

If events  $A$  and  $B$  are mutually exclusive, then

$$P(A \text{ or } B) = P(A) + P(B) .$$

More generally,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) .$$

#### The Complement Rule

$$P(\text{not } A) = 1 - P(A)$$

#### The Equally Likely Rule

If the sample space consists of  $n$  equally likely outcomes, then the probability of an event  $A$  is given by

$$P(A) = \frac{\text{number of outcomes in } A}{n} = \frac{|A|}{|S|} .$$

*Warning:* One of the most common mistakes in probability is to apply this rule when the outcomes are not equally likely.

#### Examples 1.3.1.

1. Coin Toss:  $P(\text{heads}) = \frac{1}{2}$  if heads and tails are equally likely.
2. Rolling a Die:  $P(\text{even}) = \frac{3}{6}$  if the die is fair (each of the six numbers equally likely to occur).
3. Sum of two Dice: the sum is a number between 2 and 12, but these numbers are NOT equally likely.  
There are 36 equally likely combinations of two dice:

1,1	2,1	3,1	4,1	5,1	6,1
1,2	2,2	3,2	4,2	5,2	6,2
1,3	2,3	3,3	4,3	5,3	6,3
1,4	2,4	3,4	4,4	5,4	6,4
1,5	2,5	3,5	4,5	5,5	6,5
1,6	2,6	3,6	4,6	5,6	6,6

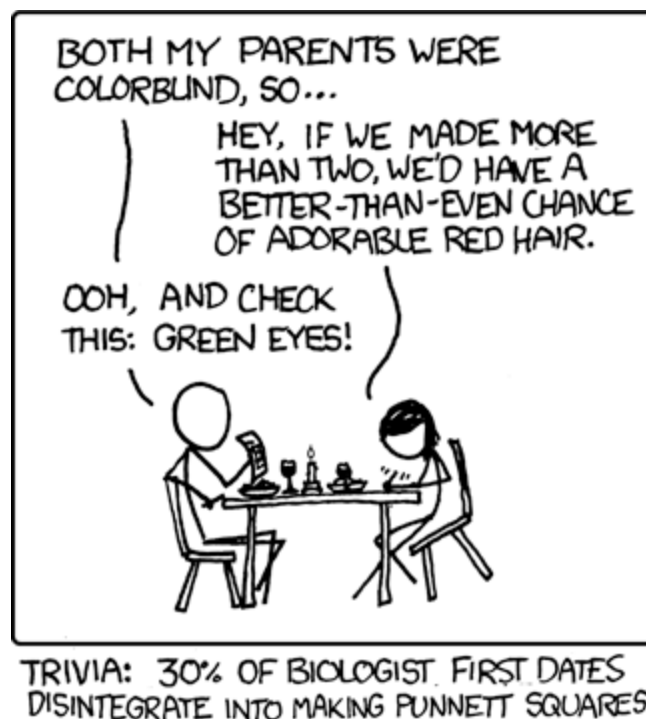
Let  $X$  be the sum of two dice.

- $P(X = 3) = \frac{2}{36} = \frac{1}{18}$
- $P(X = 7) = \frac{6}{36} = \frac{1}{6}$
- $P(\text{doubles}) = \frac{6}{36} = \frac{1}{6}$

## 4. Punnet Squares

	A	a
A	AA	Aa
a	Aa	aa

This example comes from animal or human genetics. Here, we consider a gene with two alleles: A is the dominant allele, and a is the recessive one. Each individual has two copies of every gene, so there are three possible combinations of alleles (called “genotypes”): AA, Aa, and aa. AA and Aa individuals have the dominant A physical characteristic (called the “phenotype”); aa individuals have the recessive a phenotype. Imagine that two Aa individuals mate and produce offspring. In this  $Aa \times Aa$  cross, if A is the dominant allele, then the probability of the dominant phenotype is  $\frac{3}{4}$ , and the probability of the recessive phenotype is  $\frac{1}{4}$  because each of the four possible crossings is equally likely.



Cartoon credit: <http://xkcd.com/634/>

## 1.4 Conditional Probability

**Example 1.4.1.** Q. Suppose a family has two children and one of them is a boy. What is the probability that the other is a girl?

A. We'll make the simplifying assumption that boys and girls are equally likely (which is not exactly true). Under that assumption, there are four equally likely families: BB, BG, GB, and GG. But only three of these have at least one boy, and we already know our family has at least one boy, so our sample space is really  $\{BB, BG, GB\}$ . Of these, two have a girl as well as a boy. So the probability is  $2/3$  (see Figure 1.1).

GG      

GB	BG	BB
----	----	----

      probability =  $2/3$

Figure 1.1: Illustrating the sample space for Example 1.4.1.

We can also think of this in a different way. In our original sample space of four equally likely families,

$$\begin{aligned} P(\text{at least one girl}) &= 3/4, \\ P(\text{at least one girl and at least one boy}) &= 2/4, \text{ and} \\ \frac{2/4}{3/4} &= 2/3; \end{aligned}$$

so  $2/3$  of the time when there is at least one boy, there is also a girl. We will denote this probability as  $P(\text{at least one girl} \mid \text{at least one boy})$ . We'll read this as "the probability that there is at least one girl *given that* there is at least one boy". See Figure 1.2 and Definition 1.4.

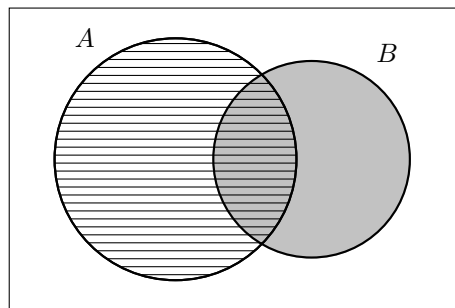


Figure 1.2: A Venn diagram illustrating the definition of conditional probability.  $P(A \mid B)$  is the ratio of the area of the football shaped region that is both shaded and striped ( $A \cap B$ ) to the area of the shaded circle ( $B$ ).

Let  $A$  and  $B$  be two events such that  $P(B) \neq 0$ . The **conditional probability** of  $A$  given  $B$  is defined by

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

If  $P(B) = 0$ , then  $P(A \mid B)$  is undefined.

**Example 1.4.2.** A class of 5th graders was asked what color should be used for the class T-shirt, red or purple. The table below contains a summary of the students' responses:

	Color	
	Red	Purple
Girls	7	9
Boys	10	8

Q. Suppose we randomly select a student from this class. Let  $R$  be the event that a child prefers a red T-shirt. Let  $B$  be the event that the child is a boy, and let  $G$  be the event that the child is a girl. Express each of the following probabilities in words and determine their values:

- $P(R)$ ,
- $P(B | R)$ ,
- $P(G | R)$ ,
- $P(R | B)$ ,
- $P(R | G)$ ,
- $P(B | G)$ .

A. The conditional probabilities can be computed in two ways. We can use the formula from the definition of conditional probability directly, or we can consider the condition event to be a new, smaller sample space and read the conditional probability from the table.

- $P(R) = 17/34 = 1/2$  because 17 of the 34 kids prefer red  
This is the probability that a randomly selected student prefers red
- $P(R | B) = \frac{10/34}{18/34} = \frac{10}{18}$  because 10 of the 18 boys prefer red  
This is the probability that a randomly selected boy prefers red
- $P(B | R) = \frac{10/34}{17/34} = \frac{10}{17}$  because 10 of the 17 students who prefer red are boys.  
This is the probability that a randomly selected student who prefers red is a boy.
- $P(R | G) = \frac{7/34}{16/34} = \frac{7}{16}$  because 7 of the 16 girls prefer red  
This is the probability that a randomly selected girl prefers red
- $P(G | R) = \frac{7/34}{17/34} = \frac{7}{17}$  because 7 of the 17 kids who prefer red are girls.  
This is the probability that a randomly selected kid who prefers red is a girl.
- $P(B | G) = \frac{0}{16/34} = 0$  because none of the girls are boys.  
This is the probability that a randomly selected girl is a boy.

One important use of conditional probability is as a tool to calculate the probability of an intersection.

Let  $A$  and  $B$  be events with non-zero probability. Then

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

This follows directly from the definition of conditional probability by a little bit of algebra and can be generalized to more than two events.

**Example 1.4.3.** Q. If you roll two standard dice, what is the probability of doubles? (Doubles is when the two numbers match.)

A. Let  $A$  be the event that we get a number between 1 and 6 on the first die. So  $P(A) = 1$ . Let  $B$  be the event that the second number matches the first. Then the probability of doubles is  $P(A \cap B) = P(A) \cdot P(B | A) = 1 \cdot \frac{1}{6} = \frac{1}{6}$  since regardless of what is rolled on the first die, 1 of the 6 possibilities for the second die will match it.

**Example 1.4.4.** Q. A 5-card hand is dealt from a standard 52-card deck. What is the probability of getting a flush (all cards the same suit)?

A. Imagine dealing the cards in order. Let  $A_i$  be the event that the  $i$ th card is the same suit as all previous cards. Then

$$\begin{aligned} P(\text{flush}) &= P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) \\ &= P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \cdot P(A_4 | A_1 \cap A_2 \cap A_3) \\ &\quad \cdot P(A_5 | A_1 \cap A_2 \cap A_3 \cap A_4) \\ &= 1 \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} \cdot \frac{9}{48} \end{aligned}$$

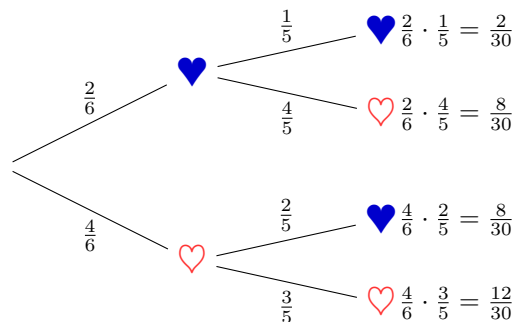
**Example 1.4.5.** Q. In a bowl are 4 red Valentine hearts and 2 blue Valentine hearts.

If you reach in without looking and select two of the Valentines, let  $X$  be the number of blue Valentines. Fill in the following probability table.

value of $X$	0	1	2
probability			

A.  $P(X = 2) = P(\text{first is blue and second is blue}) = P(\text{first is blue}) \cdot P(\text{second is blue} | \text{first is blue}) = \frac{2}{6} \cdot \frac{1}{5} = \frac{2}{30}$ . Similarly  $P(X = 0) = P(\text{first is red and second is red}) = P(\text{first is red}) \cdot P(\text{second is red} | \text{first is red}) = \frac{4}{6} \cdot \frac{3}{5} = \frac{12}{30}$ . Finally,  $P(X = 1) = 1 - P(X = 0) - P(X = 2) = 1 - \frac{14}{30} = \frac{16}{30}$

We can represent this using a **tree diagram** as well.



The edges in the tree represent conditional probabilities which we can multiply together to the probability that all events on a particular branch happen. The first level of branching represents what kind of Valentine is selected first, the second level represents the second selection.

**Example 1.4.6.** Q. Suppose a test correctly identifies diseased people 99% of the time and correctly identifies healthy people 98% of the time. Furthermore assume that in a certain population, one person in 1000 has the disease. If a random person is tested and the test comes back positive, what is the probability that the person has the disease?

A. We begin by introducing some notation. Let  $D$  be the event that a person has the disease. Let  $H$  be the event that the person is healthy. Let  $+$  be the event that the test comes back positive (meaning it indicates disease – probably a negative from the perspective of the person tested). Let  $-$  be the event that the test is negative.

- $P(D) = 0.001$ , so  $P(H) = 0.999$ .

- $P(+ | D) = 0.99$ , so  $P(- | D) = 0.01$ .

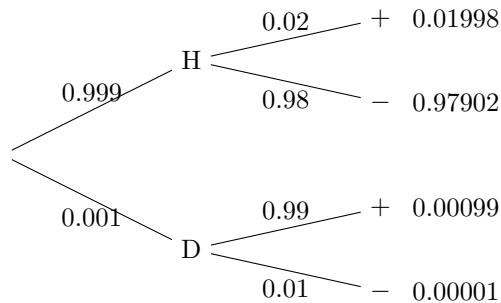
$P(+ | D)$  is called the **sensitivity** of the test. (It tells how sensitive the test is to the presence of the disease.)

- $P(- | H) = 0.98$ , so  $P(+ | H) = 0.02$ .

$P(- | H)$  is called the **specificity** of the test.

- $$\begin{aligned}
 P(D | +) &= \frac{P(D \cap +)}{P(+)} \\
 &= \frac{P(D) \cdot P(+ | D)}{P(D \cap +) + P(H \cap +)} \\
 &= \frac{0.001 \cdot 0.99}{0.001 \cdot 0.99 + 0.999 \cdot 0.02} = 0.0472.
 \end{aligned}$$

A tree diagram is a useful way to visualize these calculations.



This low probability surprises most people the first time they see it. This means that if the test result of a random person comes back positive, the probability that that person has the disease is less than 5%, even though the test is “highly accurate”. This is one reason why we do not routinely screen an entire population for a rare disease – such screening would produce many more false positives than true positives.

Of course, if a doctor orders a test, it is usually because there are some other symptoms. This changes the *a priori* probability that the patient has the disease.

### 1.4.1 Independence

Let  $A$  and  $B$  be two events such that  $P(B) = P(B | A)$ . Such events are called **independent**.

When events are independent, then  $P(A \text{ and } B) = P(A) \cdot P(B | A) = P(A) \cdot P(B)$ . This makes probability calculations much simpler – but it only applies for independent events.

**Example 1.4.7.** Q. What is the probability of rolling double sixes with standard 6-sided dice?

A. Let  $A$  be the event that the first die is a 6 and let  $B$  be the event that the second die is a 6. Since  $A$  and  $B$  are independent,  $P(A \text{ and } B) = P(A) \cdot P(B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$ .

**Example 1.4.8.** Q. What is the probability of flipping a coin five times and getting 5 heads?

A. Since each coin toss is independent of the others, the probability of getting five heads is the product of the probabilities of each coin coming up heads:

$$P(5 \text{ heads in 5 flips}) = (0.5)^5 = 0.03125$$

**Example 1.4.9.** Q. A manufacturer claims that 99% of its parts will still be functioning properly two years after purchase. If you purchase 10 of these parts, what is the probability that all 10 of them are still functioning properly two years later (assuming the manufacturer's claim is correct)?

A. Let  $G_i$  be the event that part  $i$  is still functioning properly after two years. We want to calculate

$$P(G_1 \text{ and } G_2 \text{ and } \cdots \text{ and } G_{10}) .$$

If we assume the lifetimes of the parts are independent, then

$$P(G_1 \text{ and } G_2 \text{ and } \cdots \text{ and } G_{10}) = \underbrace{.99 \cdot .99 \cdot .99 \cdots .99}_{10 \text{ of these}} = .99^{10} = 0.9043821 .$$

The independence assumption may or may not be valid. That depends on the manufacturing process. For example, if the primary way a part goes bad is that the package is dropped during shipping, then if you buy a box of 10 and the first part is bad, they will all be bad. And if the box was handled carefully and never dropped, and the first part used is good, they will likely all be good. So in that extreme case, the probability that all 10 are functioning properly after two years is 99%.



## Exercises

**1.1** Amy is a 92% free throw shooter. If she shoots 100 free throws after practice, what is the probability that she makes at least 95 of them? Use simulation to estimate this probability.

(You can use `rflip()` to simulate shooting free throws. The `prob` argument lets you set the probability. In this case, you need to set it to 0.92. Then think of a head as a made free throw and a tail as a missed free throw.)

### 1.2

- a) Use simulation to estimate the probability of rolling a difference of 2 when rolling two fair six-sided dice.
- b) Make a histogram showing the results for all of the possible differences.

**1.3** Use simulation to estimate the probability that when dealing 5 cards from a standard (well-shuffled) deck of 52 cards all five are diamonds.

You can simulate the deck of cards using the numbers 1 through 52 and consider the numbers 1 through 13 to be the diamonds. Instead of using `resample()`, which would allow you to get the same card more than once, we need to use `sample()`, which does not. (You can also use `deal()` which does the same thing.)

```
sample(1:52, 5)

## [1] 19 26 36 30 21

sample(1:52, 5)

## [1] 52 10 21 28 12

deal(1:52, 5)

## [1] 19 8 52 51 18

deal(1:52, 5)

## [1] 30 23 20 51 15
```

There is another way to make the calculation, using the function `sum()`. R can tell you how many cards are below 14 using `sum()` because R turns TRUE into 1 and FALSE into 0 when you do a sum.

```
sum(sample(1:52, 5) < 14)

## [1] 1

sum(sample(1:52, 5) < 14)

## [1] 1

sum(sample(1:52, 5) < 14)

## [1] 2
```

You can use `do()` to do this many times. (Three is *not* many. We just do a small number here for illustration purposes.)

```
do(3) * sum(sample(1:52, 5) < 14)

##      sum
## 1      0
## 2      1
## 3      1
```

**1.4** Parts in a manufacturing plant go through two quality control checks before they are shipped. 99% of parts pass inspection A and 98% parts pass inspection B. 0.5% fail both inspections.

What percentage of parts pass both inspections?

**1.5** Let  $X$  be the sum of the results of rolling two fair six-sided dice.

- What is  $P(X \text{ is even and } X < 5)$ ?
- What is  $P(X \text{ is even or } X < 5)$ ?

**1.6** Let  $Y$  be the difference between the larger and smaller number when two fair dice are rolled. (So if you roll a 2 and a 4, then the value of  $Y$  is 2.)

- What is  $P(Y = 2)$ ?
- What are the other possible values of  $Y$ ?
- Calculate the probability for each possible value of  $Y$  and put those values in a table.

**1.7** A device is assembled from two primary parts. 2% of the first type of part are defective and 3% of the other type of part are defective. The device only functions properly if both parts are functioning properly.

- a) What assumption do you need to make to calculate the probability that a device assembled in this way will function properly? Is it a reasonable assumption in this situation? Explain.
- b) What is the probability that that a device assembled in this way will function properly?

**1.8** According to the CDC, “Compared to nonsmokers, men who smoke are about 23 times more likely to develop lung cancer and women who smoke are about 13 times more likely.” According to the American Lung Association: “In 2008, 21.1 million (18.3%) women smoked in the United States compared to 24.8 million (23.1%) men.”

- a) If you learn that a person is a smoker and no nothing else about the person, what is the probability that the person is a woman?
- b) If you learn that a woman has been diagnosed with lung cancer, and you know nothing else about her, what is the probability that she is a smoker?
- c) If you learn that a man has been diagnosed with lung cancer, and you know nothing else about him, what is the probability that he is a smoker?

**1.9** A manufacturing plant has kept records that show that the number of parts produced each day and on the proportion of parts that are defective.

	Monday	Tuesday	Wednesday	Thursday
Proportion of weekly production	20%	25%	28%	27%
Rate of defective parts	2%	1.5%	1%	3%

- a) If you order a part from this company, what is the probability that it was produced on a Monday or a Thursday?
- b) If you order a part from this company and it is defective, what is the probability that it was produced on a Monday or a Thursday?
- c) If you order a part from this company and it functions properly, what is the probability that it was produced on a Monday or Thursday?

Express your answers to 3 significant digits and avoid internal rounding.



Excellent health statistics - smokers are less likely to die of age related illnesses.'

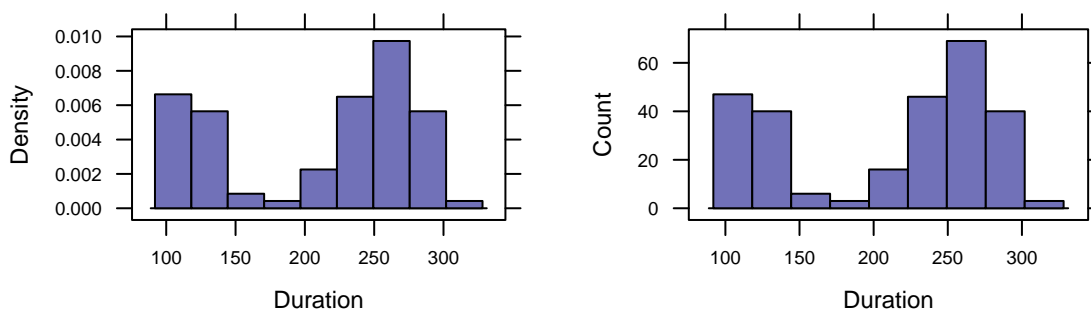
## 2

## Densities

## 2.1 Density histograms, density plots, density functions

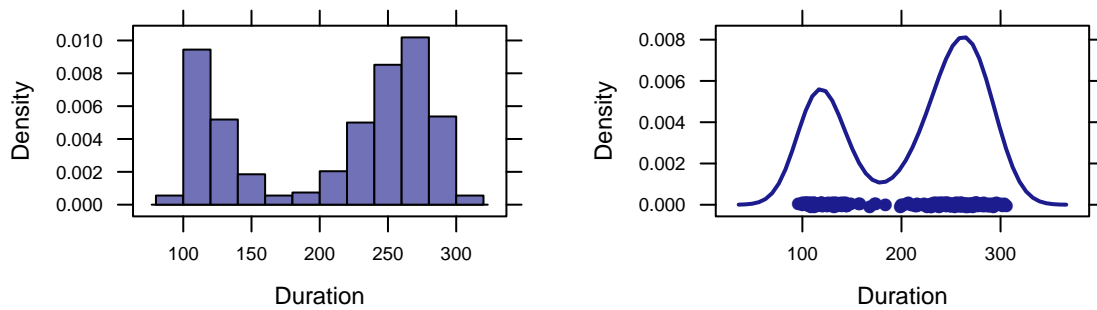
A histogram is a simple picture describing the “density” of data. Histogram bars are tall in regions where there is more data – i.e., where the data are more “dense”.

```
require(alr3)
histogram(~Duration, data = oldfaith)
histogram(~Duration, data = oldfaith, type = "count")
```



The density scale is the same scale that is used by `densityplot()`, and it is the default scale for histograms created using `histogram()` when the `mosaic` package is loaded.

```
require(alr3)
histogram(~Duration, data = oldfaith, width = 20, center = 110)
densityplot(~Duration, data = oldfaith)
```



The density scale is chosen so that the area of each rectangular bar (width times height) is equal to the proportion of the data set represented by the rectangle.

**Example 2.1.1.** Q. Use the histogram of Old Faithful eruption times to estimate the proportion of eruptions that last between 100 and 120 seconds.

A. In our histogram of Old Faithful eruption durations, the bar corresponding to the bin from 100–120 appears to have a height of about 0.09. That gives an area of 0.18 and indicates that approximately 18% of the eruptions last between 100 and 120 seconds.

```
tally(~(100 < Duration & Duration <= 120), data = oldfaith, format = "prop")

##
##      TRUE      FALSE
## 0.1888889 0.8111111
```

The key idea behind the density scale can be expressed as

$$\text{Probability} = \text{area}$$

This association of area with probability means that the total area of all the bars will always be equal to 1 if we use the density scale.

It also provides us with a way to describe a distribution with a mathematical function.

Let  $f$  be a function such that

1.  $f(x) \geq 0$  for all  $x$ ,

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

Then  $f$  is called a **density function** (or probability density function, abbreviated pdf) and describes a continuous random variable  $X$  such that

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

**Example 2.1.2.** Let  $f$  be defined by

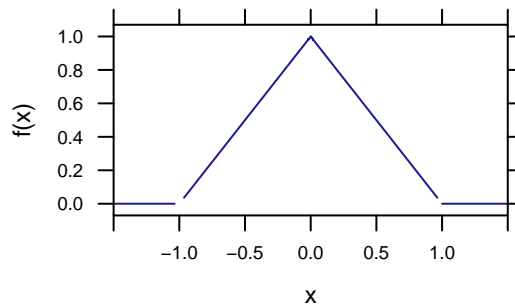
$$f(x) = \begin{cases} 1 - |x| & x \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$$

Show that  $f$  is a density function. Let  $X$  be the associated random variable, and compute the following probabilities:

1.  $P(X \leq 0)$
2.  $P(X \leq 1)$
3.  $P(X \leq \frac{1}{2})$
4.  $P(-\frac{1}{2}X \leq \frac{1}{2})$

A. While we could set up integrals for these, it is easier to solve them using geometry.<sup>1</sup>

```
f <- makeFun((1 - abs(x)) * (abs(x) <= 1) ~ x)
plotFun(f(x) ~ x, x.lim = c(-1.5, 1.5))
```



The entire area under the curve can be found as the area of a triangle with base 2 and height 1.

$$\int_{-\infty}^{\infty} f(x) dx = \int_{-1}^1 f(x) dx = \frac{1}{2} \cdot 2 \cdot 1 = 1$$

This implies that  $f$  is a density function.

1.  $P(X \leq 1) = \int_{-\infty}^1 f(x) dx = \int_{-1}^1 f(x) dx = 1$
2.  $P(X \leq \frac{1}{2}) = \int_{-\infty}^{1/2} f(x) dx = \int_{-1}^{1/2} f(x) dx = 1 - \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{7}{8}$
3.  $P(-\frac{1}{2} \leq X \leq \frac{1}{2}) = \int_{-1/2}^{1/2} f(x) dx = 1 - \frac{2}{8} = \frac{3}{4}$

We can also let R do (numerical) integration for us. There are two ways to do this. The first method uses the `integrate()` function.

```
integrate(f, -Inf, 1)

## 1 with absolute error < 9.2e-05

# this will be more accurate since we aren't asking R to approximate something that we
# already know is exactly 0
integrate(f, -1, 1)
```

<sup>1</sup>R cleverly turns TRUE and FALSE into 1 and 0 when you use them in arithmetic expressions. The definition of `f()` makes use of this conversion to simplify specifying the cases.

```
## 1 with absolute error < 1.1e-14

integrate(f, -0.5, 0.5)

## 0.75 with absolute error < 8.3e-15

# if you just want the value without the text saying how accurate the approximation is
integrate(f, -0.5, 0.5)$value

## [1] 0.75
```

An alternative approach uses `antiD()` from the `mosaic` package.

```
F <- antiD( f(x) ~ x)
F(1) - F(-1)      # total probability -- better be 1

## [1] 1

F(.5) - F(-1)     # P( -1 <= X <= 0.5 )

## [1] 0.875

F(.5) - F(-.5)    # P( -.5 <= X <= .5 )

## [1] 0.75
```

If we help R choose the anti-derivative, we get a useful function called the **cumulative distribution function**, abbreviated cdf.

If  $X$  is a random variable, then the **cumulative distribution function** (cdf) for  $X$ , often denoted  $F_X$ , is the function defined by

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

That is, the output of the cdf reports the probability of being below a particular value. The derivative of the cdf is the pdf.

**Example 2.1.3.** Continuing with our previous example, if we choose -1 as our lower endpoint, then the anti-derivative will be the cdf.

```
F <- antiD( f(x) ~ x, lower.bound = -1) # We can use -1 instead of -Inf here.
F(-1)      # this should be 0 since we chose -1 as the lower bound.

## [1] 0

F(1)      # P(X <= 1); should be 1
```



```
## [1] 1

F(.5)           # P(X <= 0.5)

## [1] 0.875

F(.5) - F(-.5)  # P( -0.5 <= X <= 0.5 )

## [1] 0.75
```

## 2.2 Working with Probability Density Functions

We have already seen that we can use a pdf  $f$  to calculate probabilities via integration, and that there is a special anti-derivative of  $f$  called the cdf such that the cdf  $F$  satisfies

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

This function can also be used to compute probabilities, since

$$P(a \leq X \leq b) = \int_a^b f(x) dx = F(b) - F(a)$$

Indeed, once we learn how to get the cdf function in R this will be our primary way to calculate probabilities in applications.

### 2.2.1 Kernels

The **kernel** of a random variable is a function that is a constant multiple of the pdf. The reason that these are interesting is that any kernel can be converted into a pdf by dividing by the appropriate constant. In particular, if

$$\int_{-\infty}^{\infty} k(x) dx = A ,$$

then  $k$  is the kernel of a random variable with pdf

$$f(x) = \frac{k(x)}{A} .$$

**Example 2.2.1.** Q. The kernel of a random variable is given by

$$k(x) = x^2 \mathbb{I}[x \in [0, 2]] .$$

Determine the pdf.

A. First we determine the value of the integral

$$\int_{-\infty}^{\infty} k(x) dx .$$

```

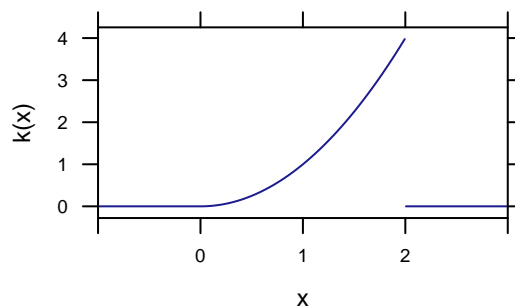
k <- makeFun(x^2 * (0 <= x & x <= 2) ~ x)
plotFun(k(x) ~ x, xlim = c(-1, 3))
integrate(k, 0, 2)

## 2.666667 with absolute error < 3e-14

K <- antiD(k(x) ~ x, lower.bound = 0)
K(2)

## [1] 2.666667

```



Since the total area is  $8/3$ , if  $\frac{k(x)}{8/3}$  is the pdf.

### 2.2.2 The mean of a continuous random variable

The definition for the mean of a continuous random variable will be motivated by the calculation of a mean of some data. **Example 2.2.2.** Q. Suppose a student has taken 10 courses and received 5 A's, 4 B's, and 1 C. Using the traditional numerical scale where an A is worth 4, a B is worth 3, and a C is worth 2, what is this student's GPA (grade point average)?

A. The first thing to notice is that  $\frac{4+3+2}{3} = 3$  is *not* correct. We cannot simply add up the values and divide by the number of values. Clearly this student should have a GPA that is higher than 3.0, since there were more A's than C's.

Consider now a correct way to do this calculation:

$$\begin{aligned}
 \text{GPA} &= \frac{4 + 4 + 4 + 4 + 4 + 3 + 3 + 3 + 3 + 2}{10} \\
 &= \frac{5 \cdot 4 + 4 \cdot 3 + 1 \cdot 2}{10} \\
 &= \frac{5}{10} \cdot 4 + \frac{4}{10} \cdot 3 + \frac{1}{10} \cdot 2 \\
 &= 4 \cdot \frac{5}{10} + 3 \cdot \frac{4}{10} + 2 \cdot \frac{1}{10} \\
 &= 3.4.
 \end{aligned}$$

The key idea here is that the mean is a **sum of values times probabilities**.

$$\text{mean} = \sum \text{value} \cdot \text{probability}$$

When working with a continuous random variable, we replace the sum with an integral and replace the probabilities with our density function to get the following definition:

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

If you recall doing center of mass problems you may recognize this integral as the first moment. (For pdfs, we don't need to divide by the "mass" because the total "mass" is the area under the curve, which will always be 1 for a random variable).

Note: It is possible that the integral used to define the mean will fail to converge. In that case, we say that the random variable has no mean or that the mean fails to exist.<sup>2</sup>

**Example 2.2.3.** Q. Compute the mean of our triangle distribution from Example 2.1.2.

A. We simply compute the integral from the definition.

$$\begin{aligned} E(X) &= \int_{-1}^1 xf(x) dx \\ &= \int_{-1}^0 x(x-1) dx + \int_0^1 x(1-x) dx \\ &= \int_{-1}^0 (x^2 - x) dx + \int_0^1 (x - x^2) dx \\ &= \left. \frac{x^3}{3} - \frac{x^2}{2} \right|_{-1}^0 + \left. \frac{x^2}{2} - \frac{x^3}{3} \right|_0^1 \\ &= \frac{1}{3} - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} = 0 \end{aligned}$$

This isn't surprising, by symmetry we would expect this result.

We could also calculate this numerically in R:

```
f <- makeFun((1 - abs(x)) * (abs(x) <= 1) ~ x)
xf <- makeFun(x * f(x) ~ x)
integrate(xf, -1, 1)

## 0 with absolute error < 3.7e-15

F <- antiD(x * f(x) ~ x, lower.bound = -1)
F(-1) # should be 0

## [1] 0

F(1)

## [1] 0
```

<sup>2</sup>Actually, we will require that  $\int_{-\infty}^{\infty} |x|f(x) dx$  converges. If this integral fails to converge, we will also say that the distribution has no mean.

### 2.2.3 The variance of a continuous random variable

Arguing similarly, we can compute the variance of a continuous random variable using

$$\text{Var}(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx$$

Note: It is possible that the integral used to define the variance will fail to converge. In that case, we say that the random variable has no variance or that the variance fails to exist.<sup>3</sup>

**Example 2.2.4.** Q. Compute the variance of the triangle random variable from the Example 2.1.2.

A.

```
f <- makeFun((1 - abs(x)) * (abs(x) <= 1) ~ x)
xxf <- makeFun((x - 0)^2 * f(x) ~ x)
integrate(xxf, -1, 1)

## 0.1666667 with absolute error < 1.9e-15

G <- antiD((x - 0)^2 * f(x) ~ x)
G(1) - G(-1)

## [1] 0.1666667
```

Some simple algebraic manipulations of the integral above shows that

$$\text{Var}(X) = E(X^2) - E(X)^2 \quad (2.1)$$

**Example 2.2.5.** Q. Compute the mean and variance of the random variable with pdf given by

$$g(x) = \frac{3x^2}{8} \mathbb{I}[x \in [0, 2]] .$$

This is the pdf computed in Example 2.2.1.

A.

```
g <- makeFun((3 * x^2/8) * (0 <= x & x <= 2) ~ x)
m <- antiD(x * g(x) ~ x, lower.bound = 0)(2) # all in one step instead of defining F or G
m

## [1] 1.5

v <- antiD((x - m)^2 * g(x) ~ x, m = m, lower.bound = 0)(2)
v

## [1] 0.15
```

<sup>3</sup>Actually, we will require that  $\int_{-\infty}^{\infty} |x|^2 f(x) dx$  converges. If this integral fails to converge, we will say that the distribution has no variance.

```
# here's the alternate computation
antiD(x^2 * g(x) ~ x, lower.bound = 0)(2) - m^2

## [1] 0.15
```

As with data, the standard deviation is the square root of the variance.

### 2.2.4 Quantiles

Quantiles solve equations of the form

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

where  $q$  is known and  $x$  is unknown. So the 50th percentile (which is the 0.5-quantile or the median) is the number such that

$$P(X \leq x) = 0.5 .$$

**Example 2.2.6.** Q. What is the 25th percentile of the triangle distribution in Example 2.1.2?

A. We need to solve for  $x$  in the following equation:

$$0.25 = P(X \leq x) .$$

We can do this by working out the integral involved:

$$\begin{aligned} 0.25 &= \int_{-1}^x 1 - |t| dt \\ &= \int_{-1}^x 1 + t dt \\ &= t + t^2/2 \Big|_{-1}^x \\ &= x + x^2/2 + 1 - 1^2/2 \\ &= x + x^2/2 + 1/2 \\ 0 &= x^2/2 + x + 1/4 \\ 0 &= 2x^2 + 4x + 1 \end{aligned}$$

So by the quadratic formula,  $x = \frac{1}{2}\sqrt{2} - 1 = -0.2928932$ .

We can check this by evaluating the cdf.

```
x <- 1/2 * sqrt(2) - 1
F(x)

## [1] 0
```

This could also be done geometrically by solving  $\frac{1}{2}y^2 = \frac{1}{4}$  and letting  $x = -1 + y$ .

## 2.3 Some Important Families of Distributions

For now, we will consider only distributions of continuous random variables (probability density functions). We will leave set aside discrete random variables (probability mass function) until quite a bit later in the course.

A family of distributions is a collection of distributions that share some common features. Typically, these are described by giving a pdf that has one or more **parameters**. A parameter is simply a number that describes (a feature of) a distribution that distinguishes between members of the family. In this section we describe briefly some of the important distributions and how to work with them in R.

### 2.3.1 Triangle Distributions

The example distribution in the previous section is usually referred to as a triangle distribution (or triangular distribution) because of the shape of its pdf. There are, of course, many triangle distributions. A triangle distribution is specified with three numbers:  $a$ , the minimum;  $b$ , the maximum, and  $c$ , the location of the peak. A triangle distribution is symmetric if the peak is halfway between the minimum and maximum ( $c = \frac{a+b}{2}$ ).

When  $X$  is a random variable with a triangle distribution, we will write  $X \sim \text{Triangle}(a, b, c)$ . For many of the most common distributions, R has several functions that facilitate computation with those distributions. The triangle distributions are not in the base R distribution, but they can be added by requiring the **triangle** package.

For each distribution, there are four functions in R that always start with a single letter followed by a name for the distribution. In the case of the triangle distributions, these functions are

Function	What it does
<code>dtriangle(x,a,b,c)</code>	Computes value of the pdf at $x$
<code>ptriangle(q,a,b,c)</code>	Computes value of the cdf at $x$ , i.e., $P(X \leq q)$
<code>qtriangle(p,a,b,c)</code>	Computes quantiles, that is a value $q$ so that $P(X \leq q) = p$
<code>rtriangle(n,a,b,c)</code>	Randomly samples $n$ values from the $\text{Triangle}(a, b, c)$ distribution

**Example 2.3.1.** Q. Let  $X \sim \text{Triangle}(0, 4, 1)$ . Use R to answer the following questions.

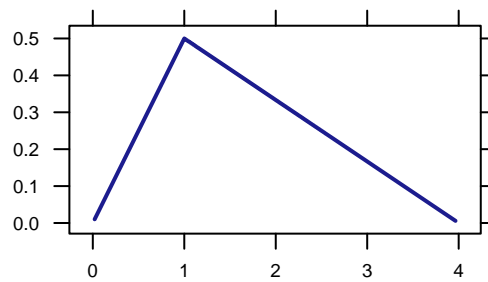
1. Plot the pdf for  $X$ .
2. What is  $P(X \leq 1)$ ?
3. What is  $P(X \leq 2)$ ?
4. What is the median of  $X$ ?
5. What is the mean of  $X$ ?

A. The `plotDist()` function in the **mosaic** package allows us to graph the pdf for any function R knows how to work with in the standard way. For example, here is a plot of the pdf of a  $\text{Triangle}(0, 4, 1)$ -distribution.

```
require(triangle) # a package that knows about triangle distributions

## Loading required package: triangle

plotDist("triangle", params = list(a = 0, b = 4, c = 1))
```



Here is the R code to answer the remaining questions.

```
ptriangle(1, 0, 4, 1) # P(X <= 4); notice that this is NOT 1/2

## [1] 0.25

ptriangle(2, 0, 4, 1) # P(X <= 4); also NOT 1/2

## [1] 0.6666667

qtriangle(0.5, 0, 4, 1) # median is the 0.5-quantile

## [1] 1.55051

T <- antiD(x * dtriangle(x, 0, 4, 1) ~ x, lower.bound = 0)
T(4) # mean of X

## [1] 1.666667

integrate(makeFun(x * dtriangle(x, 0, 4, 1) ~ x), 0, 4)

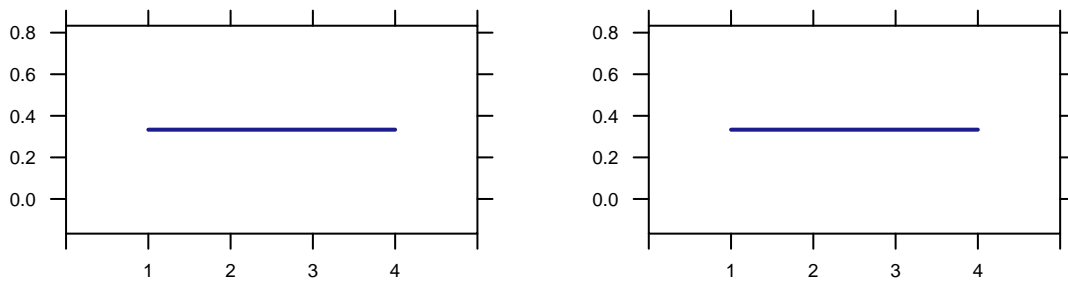
## 1.666667 with absolute error < 1.9e-14
```

### 2.3.2 Uniform Distributions

A uniform distribution is described by a constant function over some interval. Its shape is a rectangle. This makes it particularly easy to calculate probabilities for a uniform distribution. Despite its simplicity, the family of uniform distributions has many applications.

We will let  $X \sim \text{Unif}(a, b)$  denote that  $X$  is a uniform random variable on the interval from  $a$  to  $b$ . In R the parameters  $a$  and  $b$  are given more meaningful names: **min** and **max**. We can use the following code to graph the  $\text{Unif}(1, 4)$  distribution.

```
plotDist("unif", params = list(min = 1, max = 4), xlim = c(0, 5)) # using parameter names
plotDist("unif", params = list(1, 4), xlim = c(0, 5)) # without names
```



Notice that the width of the non-zero portion of the pdf is 3, so the height must be  $1/3$ .

Probabilities involving uniform distributions are easily calculated using simple geometry, but R also provides several functions for working with uniform probability distributions.

Function	What it does
<code>dunif(x,min,max)</code>	Computes value of the pdf at $x$
<code>punif(x,min,max)</code>	Computes value of the cdf at $x$ , i.e., $P(X \leq x)$
<code>qunif(p,min,max)</code>	Computes quantiles, that is a value of $x$ so that $P(X \leq x) = p$
<code>runif(n,min,max)</code>	Randomly samples $n$ values from the $\text{Unif}(\text{min}, \text{max})$ distribution

Notice the pattern to these names. They start with the same letters as the functions for the triangle distributions, but replace **triangle** with **unif**. *There are similar functions for all of the distributions in this chapter.*

**Example 2.3.2.** Q. Let  $X \sim \text{Unif}(1, 4)$ . Use R to calculate the following values and check the values using geometry:

1.  $P(X \leq 2)$
2. the 80th percentile of the distribution

A.

```
punif(2, 1, 4) # P(X <= 2 )

## [1] 0.3333333

(2 - 1) * 1/3 # P(X <= 2 ) using area

## [1] 0.3333333

qunif(0.8, 1, 4) # 80th percentile

## [1] 3.4
```

We could also get the 80th percentile by solving the equation  $\frac{1}{3}(x - 1) = 0.8$ . From this we get  $\frac{x}{3} = 0.8 + 1/3$ , so  $x = 3(0.8 + 1/3) = 2.4 + 1 = 3.4$ .



### 2.3.3 Exponential Distributions

The exponential distributions are useful for modeling the time until some “event” occurs. The model is based on the assumptions that

1. The probability of an event occurring in any small interval of time is proportional to the length of the time interval. The constant of proportionality is the rate parameter, usually denoted by  $\lambda$ .
2. The probabilities of events occurring in two small non-overlapping intervals are independent.

**Examples 2.3.3.** Here are some situations that might be well modeled by an exponential distribution:

1. The time until the next radioactive decay event is detected on a Geiger counter
2. The time until a space satellite is struck by a meteor (or some other space junk) and disabled.

The model would be good if (over some time span of interest) the chances of getting struck are always the same. It would not be such a good model if the satellite moves through time periods of relatively higher and then relatively lower chances of being struck (perhaps because we pass through regions of more or less space debris at different times of the year.)

3. The lifetime of some manufactured device.

This is a pretty simple model (we’ll learn better ones later) and most often is *too* simple to describe the interesting features of the lifetime of a device. In this model, failure is due to some external thing “happening to” the device; the device itself does not wear (or improve) over time.

We will let  $X \sim \text{Exp}(\lambda)$  denote that  $X$  has an exponential distribution with rate parameter  $\lambda$ . The kernel of such a distribution is

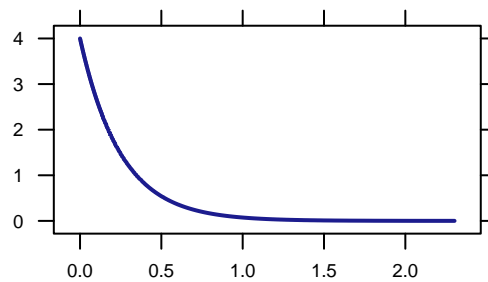
$$k(x; \lambda) = e^{-\lambda x} \mathbb{I}[x \geq 0]$$

Notice that the function describing this distribution is defined only for  $x$ -values that are real numbers greater than or equal to zero (in mathematical notation, the interval  $[0, \infty)$ .) This interval is sometimes called the “support” of the distribution. When using probability distributions to model data, it’s important to think about whether the support of the distribution matches well with the range of possible values observed in the data.

The exponential distribution function is a pretty easy function to integrate, but R provides the now familiar functions to make things even easier.

Function	What it does
<code>dexp(x,rate)</code>	Computes value of the pdf at $x$
<code>pexp(q,rate)</code>	Computes value of the cdf at $x$ , i.e., $P(X \leq q)$
<code>qexp(p,rate)</code>	Computes quantiles, that is a value $q$ so that $P(X \leq q) = p$
<code>rexp(n,rate)</code>	Randomly samples $n$ values from the $\text{Exp}(\lambda)$ distribution

```
plotDist("exp", params = list(rate = 4))
```



### 2.3.4 Gamma and Weibull Distributions

The Gamma and Weibull families of distributions are generalizations of the exponential distribution. Each family has two parameters, a rate parameter as in the exponential distribution, and an additional parameter called the shape parameter (denoted by  $\alpha$  below). The reciprocal of the rate parameter is called the scale parameter. For the Gamma distribution, R lets us use either rate or scale (and the default is rate). For the Weibull, we must use the scale.

distribution	kernel
Gamma( $\alpha, \lambda$ )	$k(x) = x^{\alpha-1} e^{-\lambda x} \mathbb{I}[x \geq 0]$
Weibull( $\alpha, \lambda$ )	$k(x) = x^{\alpha-1} e^{-\lambda x^\alpha} \mathbb{I}[x \geq 0]$

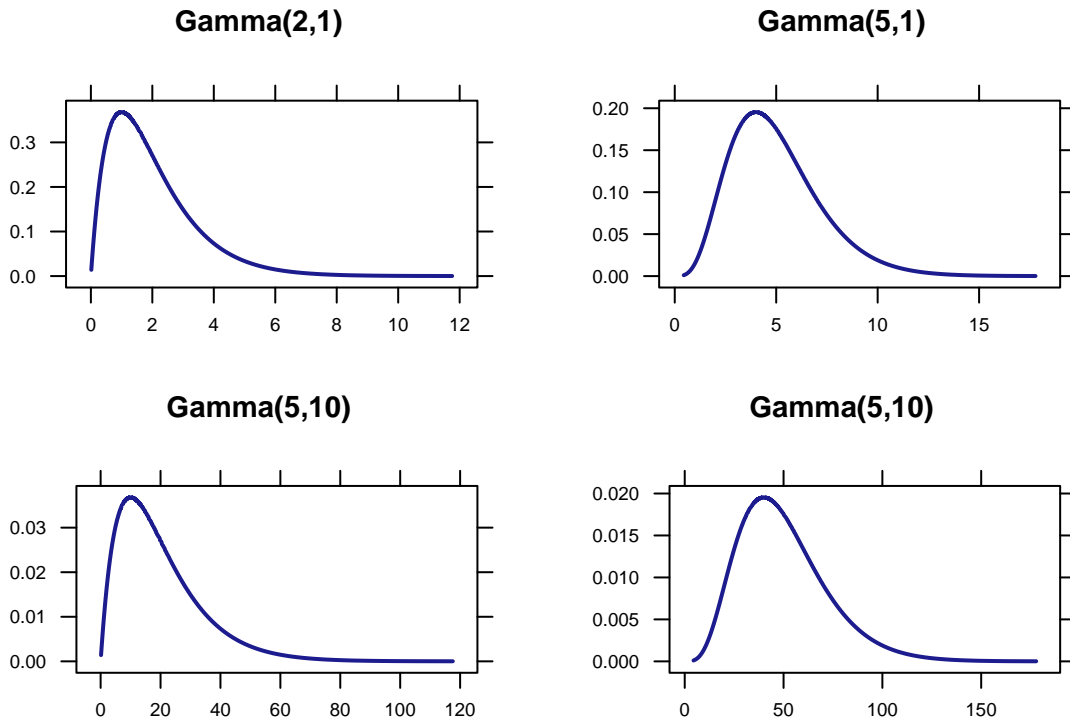
Both families of distributions are supported on the interval  $[0, \infty)$ . For the most part, we won't use these formulas in calculations, preferring to let R do the work for us. However, notice that each of these distributions has a pdf that allows for relatively simple integration. For the Gamma distributions, we need to use integration by parts ( $\alpha - 1$  times). For the Weibull distributions we can use a substitution:  $u = x^\alpha$ . In each case, when  $\alpha = 1$  we get an exponential distribution.

The now familiar functions are available for each of these distributions.

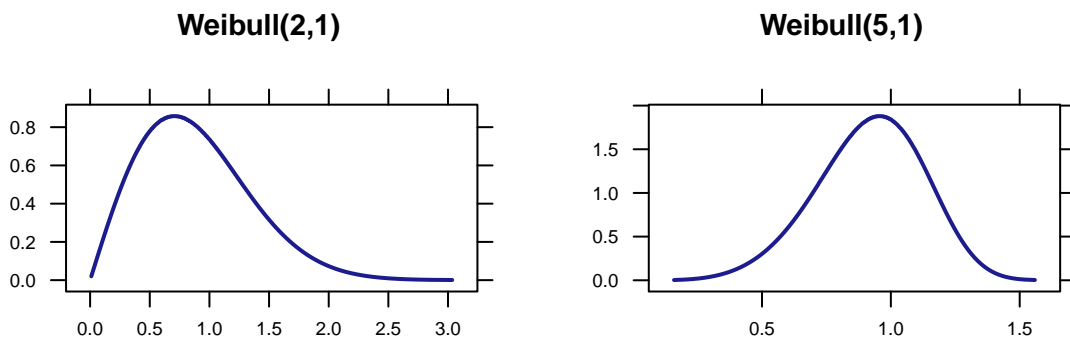
Function	What it does
<code>dgamma(x, shape, rate, scale=1/rate)</code>	Computes value of the pdf at $x$
<code>pgamma(q, shape, rate, scale=1/rate)</code>	Computes value of the cdf at $x$ , i.e., $P(X \leq q)$
<code>qgamma(p, shape, rate, scale=1/rate)</code>	Computes quantiles, that is a value $q$ so that $P(X \leq q) = p$
<code>rgamma(n, shape, rate, scale=1/rate)</code>	Randomly samples $n$ values from a Gamma distribution.
<code>dweibull(x, shape, scale=1/rate)</code>	Computes value of the pdf at $x$
<code>pweibull(q, shape, scale)</code>	Computes value of the cdf at $x$ , i.e., $P(X \leq q)$
<code>qweibull(p, shape, scale)</code>	Computes quantiles, that is a value $q$ so that $P(X \leq q) = p$
<code>rweibull(n, shape, scale)</code>	Randomly samples $n$ values from a Weibull distribution.

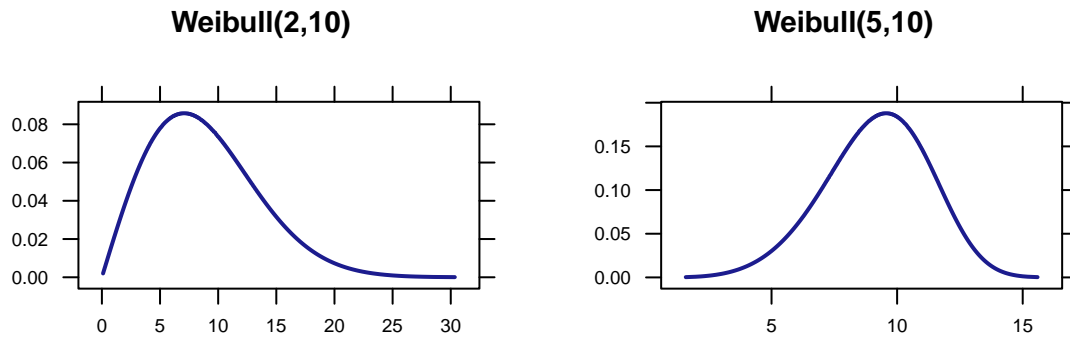
Like the exponential distributions, these distributions are skewed and only take on positive values. These distributions arise in many applications, including as more general models for lifetime. As the pictures below indicate, the shape and scale parameters are aptly named.

```
plotDist("gamma", params = list(shape = 2, rate = 1), main = "Gamma(2,1)")
plotDist("gamma", params = list(shape = 5, rate = 1), main = "Gamma(5,1)")
plotDist("gamma", params = list(shape = 2, scale = 10), main = "Gamma(5,10)")
plotDist("gamma", params = list(shape = 5, scale = 10), main = "Gamma(5,10)")
```



```
plotDist("weibull", params = list(shape = 2, scale = 1), main = "Weibull(2,1)")
plotDist("weibull", params = list(shape = 5, scale = 1), main = "Weibull(5,1)")
plotDist("weibull", params = list(shape = 2, scale = 10), main = "Weibull(2,10)")
plotDist("weibull", params = list(shape = 5, scale = 10), main = "Weibull(5,10)")
```

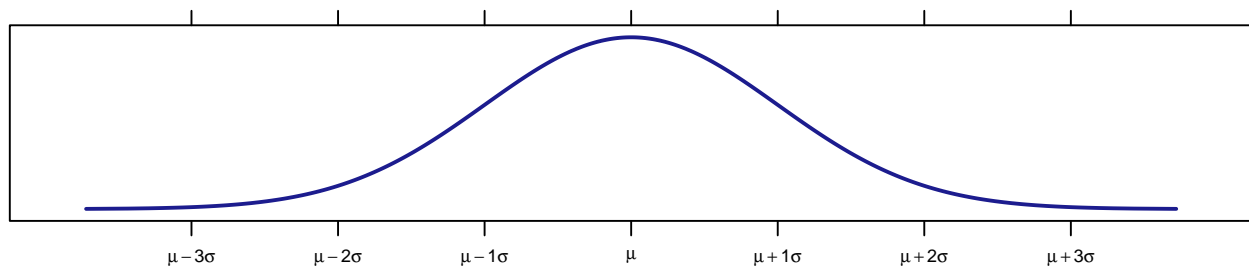




### 2.3.5 Normal Distributions

We come now to the most famous family of distributions – the normal distributions (also called Gaussian distributions). These symmetric distributions have the famous “bell shape” and are described by two parameters, the mean  $\mu$  and the standard deviation  $\sigma$ . The pdf for a  $\text{Norm}(\mu, \sigma)$  distribution is

distribution	pdf
$\text{Norm}(\mu, \sigma)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$



The inflection points of the normal distributions are always at  $\mu - \sigma$  and  $\mu + \sigma$ .

Among the normal distributions is one special distribution – the **standard normal distribution** – which has mean 0 and standard deviation 1. All other normal distributions are simply linear transformations of the standard normal distribution. That is, If  $Z \sim \text{Norm}(0, 1)$  and  $Y = a + bX$ , then  $Y \sim \text{Norm}(a, b)$ . Conversely, if  $Y \sim \text{Norm}(\mu, \sigma)$ , then  $Z = \frac{Y - \mu}{\sigma} \sim \text{Norm}(0, 1)$ .

As with the other distributions we have encountered, we have four functions that allow us to work with normal distributions in R:

Function	What it does
<code>dnorm(x, mean, sd)</code>	Computes value of the pdf at $x$
<code>pnorm(q, mean, sd)</code>	Computes value of the cdf at $x$ , i.e., $P(X \leq q)$
<code>qnorm(p, mean, sd)</code>	Computes quantiles, that is a value $q$ so that $P(X \leq q) = p$
<code>rnorm(n, mean, sd)</code>	Randomly samples $n$ values from a normal distribution.

## The 68-95-99.7 Rule

Also known as the “Empirical Rule”, the 68-95-99.7 Rule provides a set of probability benchmarks for the normal distributions because for any normal distribution:

- $\approx 68\%$  of the normal distribution is between  $\mu - \sigma$  and  $\mu + \sigma$ .
- $\approx 95\%$  of the normal distribution is between  $\mu - 2\sigma$  and  $\mu + 2\sigma$ .
- $\approx 99.7\%$  of the normal distribution is between  $\mu - 3\sigma$  and  $\mu + 3\sigma$ .

**Example 2.3.4.** Q. Before they were rescaled, SAT scores used to be approximately normally distributed with a mean of 500 and a standard deviation of 100.

1. Approximately what percent of test takers scored between 400 and 600?
2. Approximately what percent of test takers scored above 600?
3. Approximately what percent of test takers scored below 300?
4. Approximately what percent of test takers scored between 400 and 700?

A.

1. 68%
2. Since 68% are between 400 and 600, the other 32% must be outside that range, half above and half below. So 16% are above 600.
3. Since 95% are between 300 and 700, the other 5% must be outside that range, half above and half below. So 2.5% are below 300.
4. 16% are below 400 and 2.5% are above 700, so the remaining 81.5% must be between 400 and 700.

Of course, we can get more accurate results using R:

```
pnorm(600, 500, 100) - pnorm(400, 500, 100)

## [1] 0.6826895

pnorm(700, 500, 100) - pnorm(300, 500, 100)

## [1] 0.9544997

pnorm(300, 500, 100)

## [1] 0.02275013

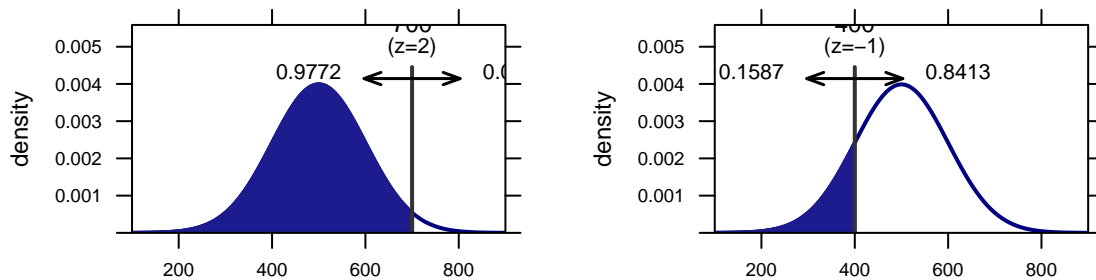
pnorm(700, 500, 100) - pnorm(400, 500, 100)

## [1] 0.8185946
```

The `xpnorm()` function will additionally draw pictures of the normal distribution with a portion of the distribution shaded in.

```
xpnorm(700, 500, 100) - xpnorm(400, 500, 100)

##
## If X ~ N(500,100), then
##
## P(X <= 700) = P(Z <= 2) = 0.9772
## P(X > 700) = P(Z > 2) = 0.0228
##
## If X ~ N(500,100), then
##
## P(X <= 400) = P(Z <= -1) = 0.1587
## P(X > 400) = P(Z > -1) = 0.8413
## [1] 0.8185946
```



**Example 2.3.5.** We can use `qnorm()` to compute percentiles. For example, let's calculate the 75th percentile for SAT distributions.

```
qnorm(0.75, 500, 100)

## [1] 567.449
```

### 2.3.6 Beta Distributions

The Beta distributions have support on the interval  $(0, 1)$ , so they can provide a model for proportions or other quantities that are bounded between 0 and 1.<sup>4</sup> The Beta distributions have two parameters, imaginatively called `shape1` and `shape2`. The kernel of the Beta distributions is a product of a power of  $x$  and a power of  $(1 - x)$ :

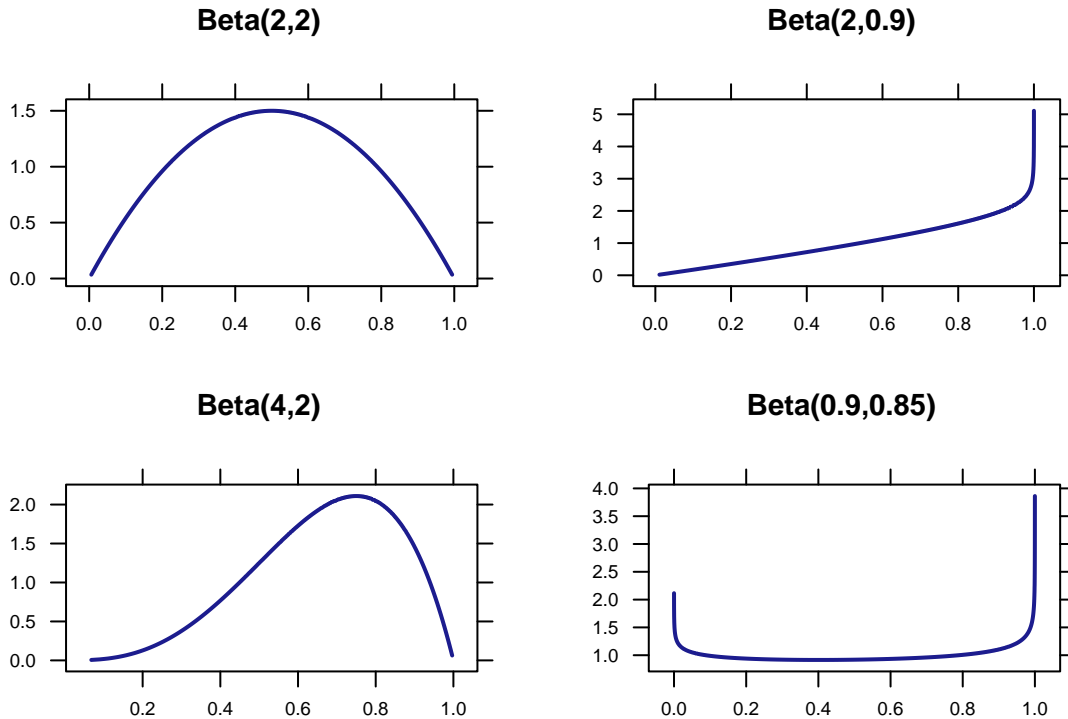
$$k(x; \alpha, \beta) = x^{\alpha-1} (1-x)^{\beta-1} \mathbb{I}[x \in [0, 1]]$$

When  $\alpha = \beta$ , the distribution is symmetric, and when  $\alpha = \beta = 1$ , we have the  $\text{Unif}(0, 1)$ -distribution.

The two shape parameters provide a wide variety of shapes.

<sup>4</sup>A more general version of the Beta distributions can do the same thing for quantities bounded by any two numbers. This more general family of distributions has four parameters.

```
plotDist("beta", params = list(shape1 = 2, shape2 = 2), main = "Beta(2,2)")
plotDist("beta", params = list(shape1 = 2, shape2 = 0.9), main = "Beta(2,0.9)")
plotDist("beta", params = list(shape1 = 4, shape2 = 2), main = "Beta(4,2)")
plotDist("beta", params = list(shape1 = 0.9, shape2 = 0.85), main = "Beta(0.9,0.85)")
```



Function	What it does
<code>dbeta(x, shape1, shape2)</code>	Computes value of the pdf at $x$
<code>pbeta(q, shape1d, shape2)</code>	Computes value of the cdf at $x$ , i.e., $P(X \leq q)$
<code>qbeta(p, shape1, shape2)</code>	Computes quantiles, that is a value $q$ so that $P(X \leq q) = p$
<code>rbeta(n, shape1, shape2)</code>	Randomly samples $n$ values from a Beta distribution.

## 2.4 Fitting Distributions to Data

Suppose we think a family of distributions would make a good model for some situation. How do we decide which member of the family to use? The simple answer is that we should choose the one that fits “best.” The trick is deciding what it means to fit well. In fact there is more than one way to measure how well a distribution fits a data set.

**Example 2.4.1.** We can use the following code to load a data set that contains three year’s worth of mean hourly wind speeds (mph) in Twin Falls, ID. This kind of data is often used to estimate how much power could be generated from a windmill placed in a given location.

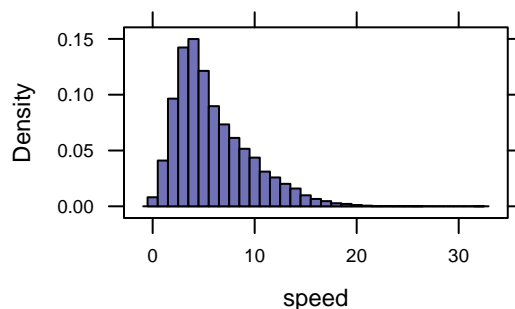
```
Wind <- read.csv("http://www.calvin.edu/~rpruim/data/stob/TwinfallsWind.csv")
head(Wind, 2)
```

```
##           date time speed
## 1 1/1/2010 0:00  2.24
## 2 1/1/2010 1:00  2.42

tail(Wind, 2)

##           date time speed
## 26272 12/31/2012 22:00  3.88
## 26273 12/31/2012 23:00  5.04

histogram(~speed, data = Wind, width = 1)
```



As we can see, the distribution is skewed, but it doesn't look like an exponential distribution would be a good fit. Of the distributions we have seen, it seems like a Weibull or Gamma distribution would be a potentially good choice. A Weibull model has often been used as a model for mean hourly wind speed, and the shape of our histogram indicates that this is a reasonable family of distributions.

Q. Which Weibull distribution is the best model for our data?

A. The `fitdistr()` in the **MASS** package uses the method of **maximum likelihood** to fit univariate (one variable) distributions.

```
fitdistr(Wind$speed, "weibull")

## Error in fitdistr(Wind$speed, "weibull"): Weibull values must be > 0
```

For `fitdistr()` to fit a Weibull distribution, all of the data must be positive, but our data includes some 0's.

```
tally(~speed == 0, data = Wind)

##
## TRUE FALSE
## 48 26225
```

Let's see how small the smallest non-zero measurements are.



```
min(~speed, data = subset(Wind, speed > 0))

## [1] 0.01
```

This may well be a simple rounding issue, since the wind speeds are recorded to the nearest 0.01 and 0.01 is the smallest positive value. Let's create a new variable that moves each value of 0 to 0.0025 and try again. Why 0.0025? If we think that 0.01 represents anything in the range 0.005 to 0.015, which would round to 0.01, then 0 represents anything in the range 0 to 0.005. 0.0025 is the middle of that range.

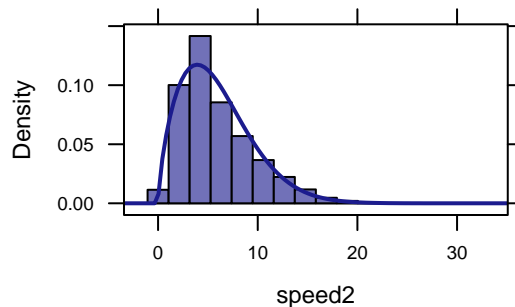
```
Wind <- transform(Wind, speed2 = ifelse(speed > 0, speed, 0.0025))
fitdistr(Wind$speed2, "weibull")

##          shape          scale
## 1.694422851 6.650586935
## (0.007957624) (0.025551827)
```

This says that the best fitting (in the sense of maximum likelihood) Weibull distribution is the Weibull(1.69, 6.65)-distribution.

The `histogram()` function has an option to overlay the distribution fit by `fitdistr()` so we can see how good the fit is graphically.

```
histogram(~speed2, data = Wind, fit = "weibull")
```

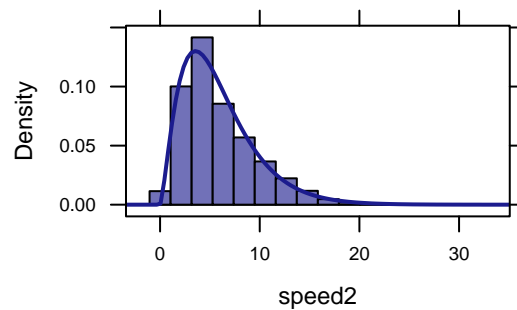


**Example 2.4.2.** As an alternative, we could fit a Gamma distribution to the wind speed data.

```
fitdistr(Wind$speed2, "gamma")

##          shape          rate
## 2.495582854 0.421178362
## (0.020485581) (0.003828652)

histogram(~speed2, data = Wind, fit = "gamma")
```



By eye, it appears that the Gamma distribution fits this data set slightly better, but there may other reasons to prefer the Weibull distribution. In fact, there has been a good deal of research done regarding which distributions to use for wind speed data fitting. The answer to the question of which distributions should be used seems to be that it depends on the purpose for your modeling: “The fact that different distributions excel under different applications motivates further research on model selection based upon the engineering parameter of interest.” [?]

**Example 2.4.3.** 1986–87 was a good season for Michael Jordan, a famous former NBA basketball player. Possible models for the points scored each game that season are normal, Weibull, and Gamma distributions. The normal distributions might be a good choice if we think that the distributions is roughly symmetric (very good games are about the same amount above average as the very poor games are below average). Weibull and Gamma distributions have the built in feature that scores cannot be negative and would allow for a skewed distribution. The `fitdistr()` function in the **MASS** package can fit each of these.

```
require(fastR) # the Jordan8687 data set is in this package
fitdistr(Jordan8687$Points, "normal")

##      mean      sd
## 37.0853659  9.8639541
## ( 1.0892915) ( 0.7702454)

fitdistr(Jordan8687$Points, "weibull")

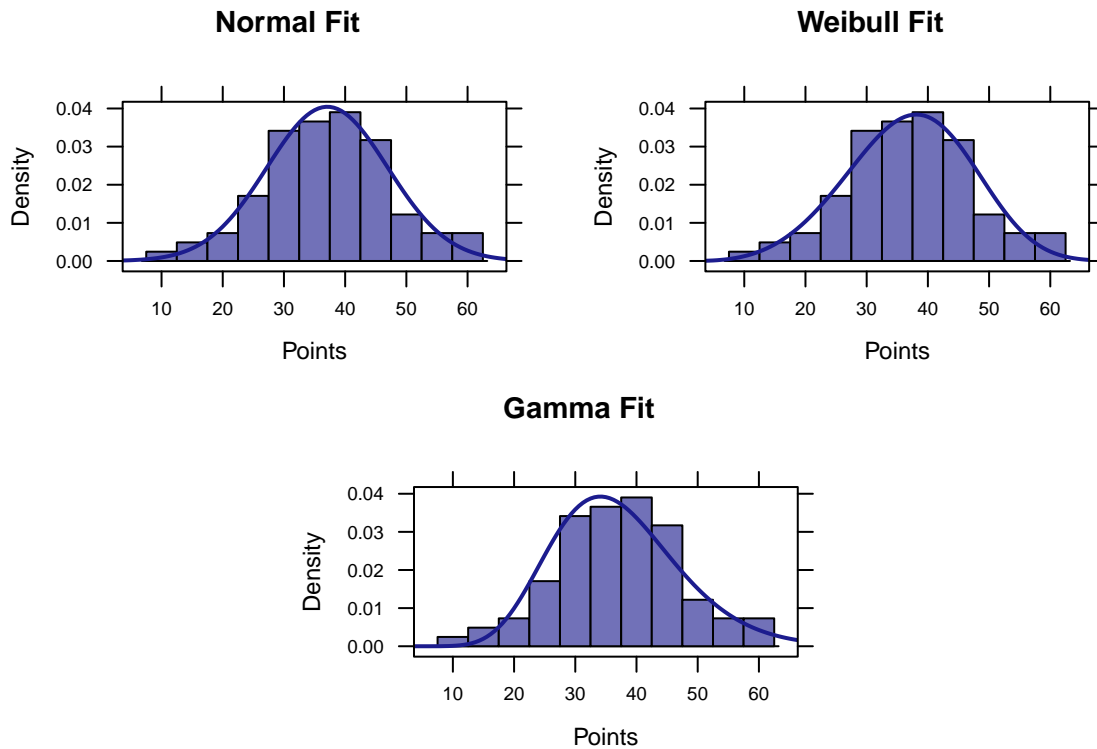
##      shape      scale
##  4.1227692  40.7746012
## ( 0.3454908) ( 1.1516943)

fitdistr(Jordan8687$Points, "gamma")

##      shape      rate
## 12.4284300  0.3351303
## ( 1.9153529) ( 0.0527028)
```

We can use a histogram with overlaid density curve to see how well these fits compare to the data.

```
histogram(~Points, Jordan8687, fit = "normal", width = 5, main = "Normal Fit")
histogram(~Points, Jordan8687, fit = "weibull", width = 5, main = "Weibull Fit")
histogram(~Points, Jordan8687, fit = "gamma", width = 5, main = "Gamma Fit")
```

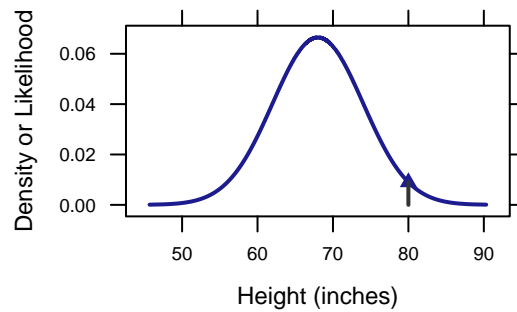


The three fits are similar, but not identical.

### 2.4.1 Maximum Likelihood

The `fitdistr()` function uses the maximum likelihood method to estimate distribution parameters. The maximum likelihood method is one of the most commonly used estimation methods in all of statistics because (1) it can be used in a wide range of applications, and (2) the resulting estimators have some desirable properties. Maximum likelihood estimation tries to choose the parameter values that *maximize* the *likelihood* of the observed data.

First, let's think about the “likelihood” of an individual observed data-point. The likelihood of the data-point is just the probability density function (or probability mass function) for the distribution of interest, evaluated at the value observed in the data. The likelihood gives some indication of how frequently we'd expect to observe this value, but it is *not* a probability (for one thing, likelihoods can exceed 1). The figure below illustrates that the likelihood of observing a person 80 inches (6 feet, 8 inches) tall, if the person comes from a population whose heights are Normally distributed with a mean of 68 inches and a standard deviation of 6 inches is about 0.009:



Given a set of specific parameter values, the likelihood of an entire observed data-set can be calculated by obtaining the value of the likelihood of each observed data-point, and summing these over all the observed data points. Then, we can find the maximum likelihood parameter estimates by trying many candidate parameter values until satisfied that we have found the ones that maximize the likelihood. (The numerical methods used are usually a bit more sophisticated than “guessing lots of random candidate values”, but we won’t get into the details here. In some cases, it is also possible to write down a mathematical expression for the likelihood of the data given the parameters, and maximize it analytically.)

We’ll illustrate the main ideas of maximum likelihood with a simple example.

**Example 2.4.4.** Michael has three dice in his pocket. One is a standard die with six sides, another has four sides, and the third has ten sides. He challenges you to a game. Without showing you which die he is using, Michael is going to roll a die 10 times and report to you how many times the resulting number is a 1 or a 2. Your challenge is to guess which die he is using.

Q. Michael reports that 3 of the 10 rolls resulted in a 1 or a 2. Which die do you think he was using?

A. The probability of obtaining a 1 or a 2 is one of  $\frac{1}{2}$ ,  $\frac{1}{3}$ , or  $\frac{1}{5}$ , depending on which die is being used. Our data are possible with any of the three dice, but let’s see how likely they are in each case.

- If  $P(\text{roll 1 or 2}) = \frac{1}{5}$ , then the probability of obtaining exactly Michael’s data is

$$\left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^7 = 0.0599323 .$$

(Whatever the order, there will be 3 events with probability  $1/5$  and 7 with probability  $4/5$ . Since the events are independent, we can multiply all of these probabilities.)

- If  $P(\text{roll 1 or 2}) = \frac{1}{3}$ , then the probability of obtaining exactly Michael’s data is

$$\left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^7 = 0.0021677 .$$

- If  $P(\text{roll 1 or 2}) = \frac{1}{2}$ , then the probability of obtaining exactly Michael’s data is

$$\left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^7 = 0.0016777 .$$

Of these, the largest likelihood is for the case that  $P(\text{roll 1 or 2}) = \frac{1}{3}$ , i.e., for the standard, six-sided die. Our data would be more likely to occur with that die than with either of the other two – it is the maximum likelihood die.

In general, maximum likelihood calculations are harder because instead of having only 3 choices, there will be infinitely many choices, and instead of having only one parameter, there may be multiple parameters. So

techniques from (multi-variable) calculus or numerical approximation methods are often used to maximize the likelihood function. The `fitdistr()` function uses pre-derived formulas for some distributions and numerical approximation methods for others. In some cases, you will get warning messages about attempts to apply a function to values that don't make sense (trying to take logs or square roots of negative numbers, zero in the denominator, etc.) as the numerical approximation algorithm explores options in an attempt to find the best fit. The help documentation for `fitdistr()` explains which distributions it can handle and what method is used for each.

## 2.4.2 The method of moments

An easy (but sometimes fairly crude) way to estimate the parameters of a distribution is the method of moments. You will often see this method used in engineering textbooks, especially if they do not rely on software that implements other methods (like the maximum likelihood method).

The basic idea is to set up a system of equations where we set the mean of the data equal to the mean of the distribution, the variance of the data equal to the variance of the distribution, etc.<sup>5</sup>

To employ this method, we need to know the means and variances of our favorite families of distributions (in terms of the parameters of the distributions). For all of the distributions we have seen, one can work out formulas for the means and variances in terms of the parameters involved. These are listed in Table 2.1

**Example 2.4.5.** Let's return to the wind speeds in Example 2.4.1. The formulas for the mean and variance of a Weibull distribution involve the gamma function  $\Gamma()$ , which might be unfamiliar to you. So let's simplify things.

Theoretical properties and observations of wind speeds at other locations suggest that using a shape parameter of  $\alpha = 2$  is often a good choice (but shape does differ from location to location depending on how consistent or variable the wind speeds are). The Weibull distributions with  $\alpha = 2$  have a special name, they are called the **Rayleigh** distributions. So  $\text{Rayleigh}(\beta) = \text{Weibull}(\alpha = 2, \beta)$ . In this case, from Table 2.1, we see that to calculate the mean we need the value of  $\Gamma(1 + \frac{1}{2}) = \Gamma(1.5) = \sqrt{\pi}/2$ .

```
gamma(1.5)

## [1] 0.8862269

sqrt(pi)/2

## [1] 0.8862269
```

From Table 2.1 we see that the mean of a  $\text{Rayleigh}(\beta)$ -distribution is

$$E(X) = \beta \frac{\sqrt{\pi}}{2}$$

Now we can choose our estimate  $\hat{\beta}$  for  $\beta$  so that

$$\hat{\beta} \frac{\sqrt{\pi}}{2} = \bar{x};$$

That is,

$$\hat{\beta} = \frac{2\bar{x}}{\sqrt{\pi}}$$

---

<sup>5</sup>If our distribution has more than 2 parameters, we will need higher moments, which we will not cover here.

```
x.bar <- mean(~speed, data = Wind)
x.bar

## [1] 5.925238

beta.hat <- x.bar * 2/sqrt(pi)
beta.hat

## [1] 6.685915
```

So our method of moments fit for the data is a  $\text{Rayleigh}(6.69) = \text{Weibull}(2, 6.69)$

Although the Rayleigh distributions are not as flexible as the Weibull or Gamma distributions, and although maximum likelihood is generally preferred over the method of moments, the method of moments fit of a Rayleigh distribution does have one advantage: it can be computed even if all you know is the mean of some sample data. Sometimes, that is all you can easily get your hands on (because the people who collected the raw data only report numerical summaries). You can find average wind speeds of for many locations online, for example here: <http://www.wrcc.dri.edu/htmlfiles/westwind.final.html>

**Example 2.4.6.** For distributions with two parameters, we solve a system of two equations with two unknowns. For the normal distributions this is particularly easy since the parameters are the mean and standard deviation, so we get

$$\begin{aligned}\hat{\mu} &= \bar{x} \\ \hat{\sigma}^2 &= s_x^2\end{aligned}$$

```
x.bar <- mean(~speed, data=Wind); x.bar

## [1] 5.925238

v <- var(~speed, data=Wind); v

## [1] 13.34635

sqrt(v)

## [1] 3.653265
```

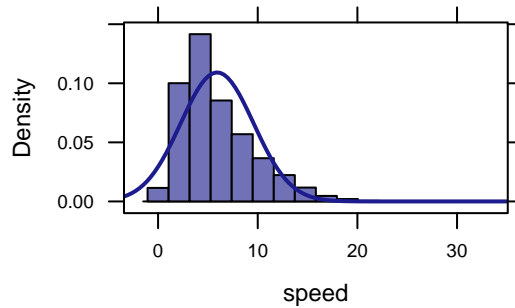
So the method of moments suggests a  $\text{Norm}(5.93, 3.65)$  distribution. In this case, the method of moments and maximum likelihood methods give the same results.

```
fitdistr(Wind$speed, "normal")

##          mean          sd
## 5.92523770 3.65319577
## (0.02253814) (0.01593687)
```

But this doesn't mean that the fit is particularly good. Indeed, a normal distribution is not a good choice for this data. We know that wind speeds can't be negative and we have other distributions (exponential, Weibull, and Gamma, for example) that are also never negative. So choosing one of those seems like a better idea. The following plot shows, as we expected, that the normal distribution is not a particularly good fit.

```
histogram(~speed, data = Wind, fit = "normal")
```



It is important to remember that the best fit using a poor choice for the family of distributions might not be a useful fit. The choice of distributions is made based on a combination of theoretical considerations, experience from previous data sets, and the quality of the fit for the data set at hand.

## 2.5 Quantile-Quantile Plots

To this point we have looked at how well a distribution fits the data by overlaying a density curve on a histogram. While this is instructive, it is not the easiest way to make a graphical comparison between a data set and a theoretical distribution. Our eyes are much better at judging whether something is linear than they are at judging whether shapes have a particular kind of curve. Furthermore, certain optical misperceptions tend to cause people to exaggerate some kinds of differences and underestimate others.

Quantile-quantile plots offer an alternative approach. As the name suggests, the idea is to compare the quantiles of our data to the quantiles of a theoretical distribution. These are then plotted as a scatter plot. Let's go through those steps with a small data set so we can see all the moving parts, then we'll learn how to automate the whole process using `qqmath()`.

### 2.5.1 Normal-Quantile Plots

The normal distributions are especially important for statistics, so normal-quantile plots will be our most important example of quantile-quantiles plots. Also, special properties of the normal distributions make normal-quantile plots especially easy and useful. We will illustrate the construction of these plots using a data set containing Michael Jordan's game by game scoring output from the 1986–87 basketball season.

**Example 2.5.1.** Let's begin by forming a randomly selected sample of 10 basketball games.

```
set.seed(123) # so you can get the same sample if you like.
SmallJordan <- sample(Jordan8687, 10)
SmallJordan

##      Game Points orig.id
```

distribution	pdf	mean	variance
Triangle: $\text{Triangle}(a, b, c)$	$\begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{if } x \in [a, c], \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{if } x \in [c, b], \\ 0 & \text{otherwise} \end{cases}$	$\frac{a+b+c}{3}$	$\frac{a^2+b^2+c^2-ab-ac-bc}{18}$
Uniform: $\text{Unif}(a, b)$	$\begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b], \\ 0 & \text{otherwise} \end{cases}$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
Standard normal: $\text{Norm}(0, 1)$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$	0	1
Normal: $\text{Norm}(\mu, \sigma)$	$\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$	$\mu$	$\sigma^2$
Exponential: $\text{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$
Gamma: $\text{Gamma}(\alpha, \lambda = \frac{1}{\beta})$	$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$	$\alpha/\lambda = \alpha\beta$	$\alpha/\lambda^2 = \alpha\beta^2$
Weibull: $\text{Weibull}(\alpha, \beta = \frac{1}{\lambda})$	$\frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha}$	$\beta\Gamma(1 + \frac{1}{\alpha})$	$\beta^2 \left[ \Gamma(1 + \frac{2}{\alpha}) - [\Gamma(1 + \frac{1}{\alpha})]^2 \right]$
Beta: $\text{Beta}(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Table 2.1: Some common continuous distributions. Standard names for parameters that appear in several distributions include **rate** ( $\lambda$ ), **shape** ( $\alpha$ ), and **scale** ( $\beta$ ). In the normal distributions,  $\mu$  and  $\sigma$  are called **mean** and **sd** in R, and in the uniform distributions,  $a$  and  $b$  are called **min** and **max**. The function  $\Gamma(x)$  that appears in the formulas for the Weibull and Beta distributions is a kind of continuous extrapolation from the factorial function. The **gamma()** function will calculate these values.



```
## 24 24 27 24
## 64 64 44 64
## 33 33 31 33
## 70 70 40 70
## 74 74 26 74
## 4 4 33 4
## 41 41 27 41
## 67 67 40 67
## 76 76 31 76
## 34 34 43 34
```

```
probs <- seq(0.05, 0.95, by=0.10)
probs
```

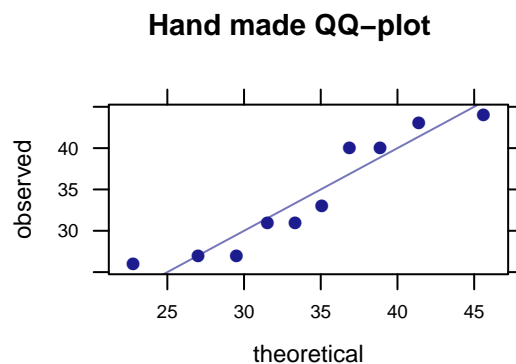
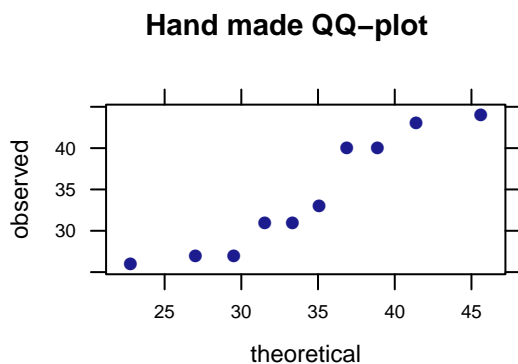
```
## [1] 0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95
```

```
observed <- sort(SmallJordan$Points) # sorted observations
theoretical <- qnorm( probs, mean=mean(observed), sd=sd(observed) ) # theoretical quantiles
QQData <- data.frame(observed=observed, theoretical=theoretical)
QQData
```

```
## observed theoretical
## 1 26 22.78304
## 2 27 27.00609
## 3 27 29.51835
## 4 31 31.52548
## 5 31 33.32778
## 6 33 35.07222
## 7 40 36.87452
## 8 40 38.88165
## 9 43 41.39391
## 10 44 45.61696
```

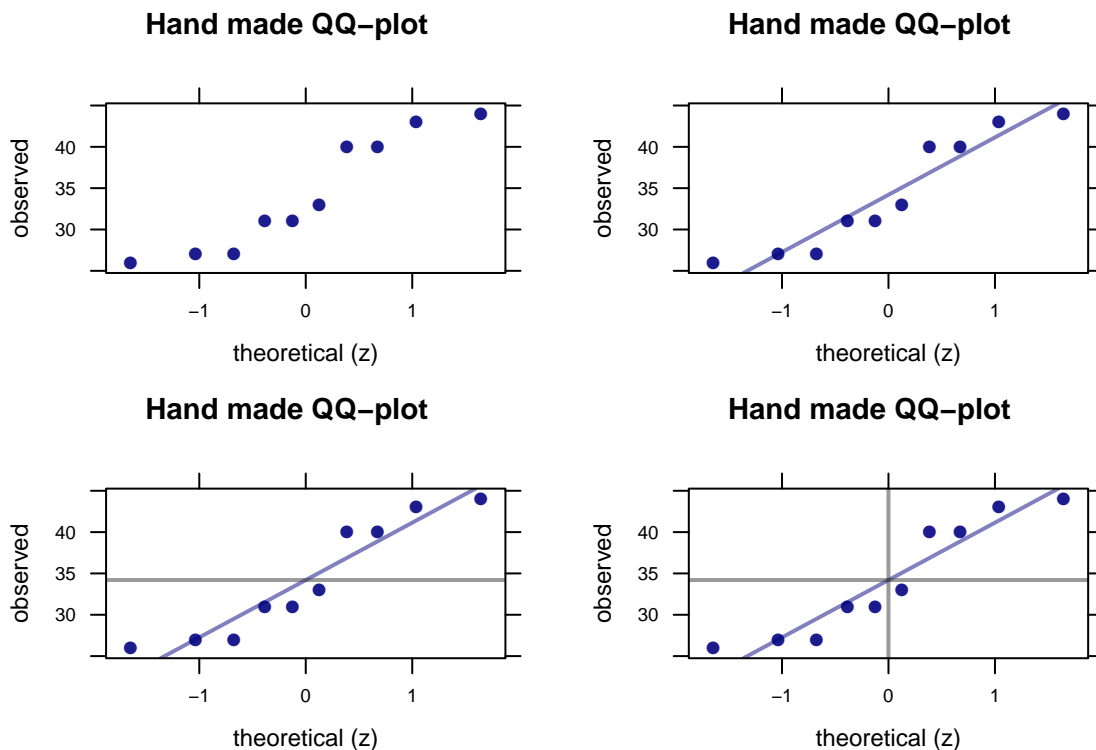
If the observed data matched the theoretical quantiles perfectly, a scatter plot would place all the points on the line with slope 1 passing through the origin.

```
xyplot(observed ~ theoretical, data = QQData, main = "Hand made QQ-plot")
plotFun(x ~ x, add = TRUE, alpha = 0.6)
```



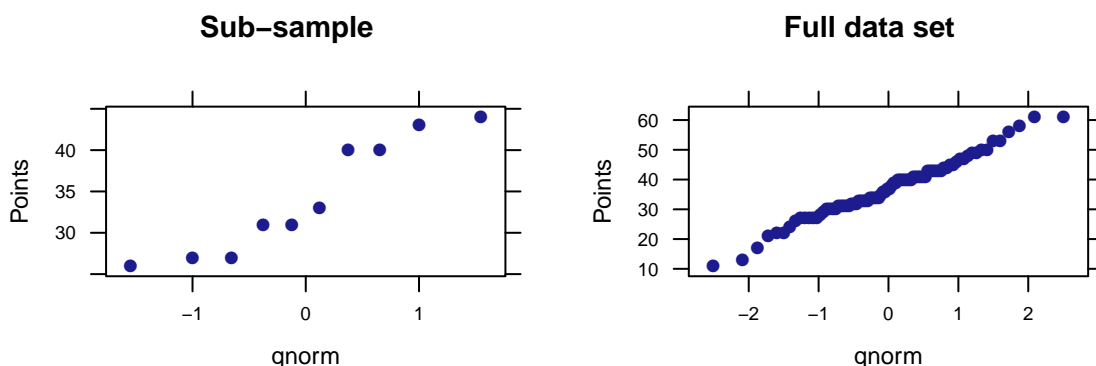
Even better, we don't need to know the mean and standard deviation in advance, because all normal distributions are linear transformations of the  $\text{Norm}(0, 1)$ -distribution. So our standard practice will be to compare our data to the  $\text{Norm}(0, 1)$ -distribution. If  $X \sim \text{Norm}(\mu, \sigma)$ , then  $X = \mu + \sigma Z$  where  $Z \sim \text{Norm}(0, 1)$ , so a plot of  $X$  vs.  $Z$  will have slope  $\sigma$  and intercept  $\mu$ .

```
theoretical2 <- qnorm(probs, mean = 0, sd = 1) # theoretical quantiles from Norm(0,1)
QQData2 <- data.frame(observed = observed, theoretical = theoretical2)
xyplot(observed ~ theoretical, data = QQData2, main = "Hand made QQ-plot", xlab = "theoretical (z)")
ladd(panel.abline(mean(SmallJordan$Points), sd(SmallJordan$Points), alpha = 0.5, col = "navy"))
ladd(panel.abline(h = mean(SmallJordan$Points), alpha = 0.5))
ladd(panel.abline(v = 0, alpha = 0.5))
```



This whole process is automated by the `qqmath()` function.

```
qqmath(~Points, SmallJordan, main = "Sub-sample")
qqmath(~Points, Jordan8687, main = "Full data set")
```

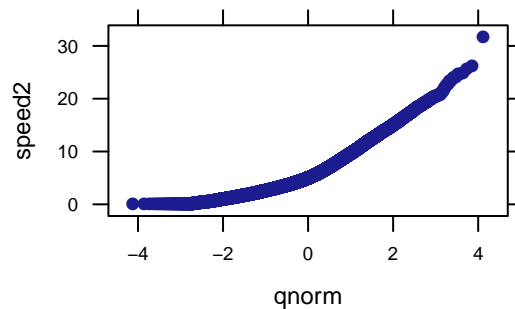


## 2.5.2 Other distributions

Working with other distributions is similar, but most families of distributions don't have a single "master example" to which we can make all comparisons, so we need to pick a particular member of the family (either by fitting or for some theoretical reason).<sup>6</sup>

**Example 2.5.2.** Let's build a quantile-quantile plot for our wind speed data comparing to normal, gamma and Weibull distributions. We can automate this, but we need to tell `qqmath()` how to calculate the quantiles.

```
qqmath(~speed2, data = Wind) # normal-quantile plot; normal is not a good model
```



The normal model does not fit well, but both Gamma and Weibull are reasonable models:

```
fitdistr(Wind$speed2, "gamma")

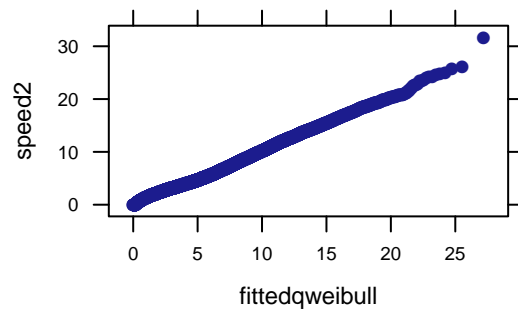
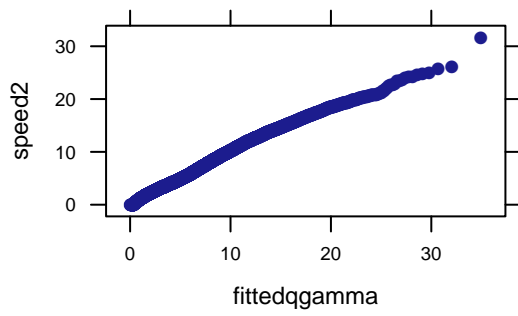
##      shape      rate
## 2.495582854 0.421178362
## (0.020485581) (0.003828652)

fitdistr(Wind$speed2, "Weibull")

##      shape      scale
## 1.694422851 6.650586935
## (0.007957624) (0.025551827)

fittedqgamma <- makeFun( qgamma(p, shape=2.496, rate=0.421) ~ p )
fittedqweibull <- makeFun( qweibull(p, shape=1.694, scale=6.651) ~ p )
qqmath( ~speed2, data=Wind, distribution=fittedqgamma )
qqmath( ~speed2, data=Wind, distribution=fittedqweibull )
```

<sup>6</sup>There are a few other families of distributions that have a prototypical member such that all other members are a linear transformation of the prototype. The exponential family is one such family.



## Exercises

**2.1** Let  $f(x) = 5/4 - x^3$  on  $[0, 1]$ .

- a) Show that  $f$  is a pdf.
- b) Calculate  $P(X \leq \frac{1}{2})$ .
- c) Calculate  $P(X \geq \frac{1}{2})$ .
- d) Calculate  $P(X = \frac{1}{2})$ .

**2.2** Repeat parts (2) – (4) of Example 2.3.1 using geometry rather than R.

**2.3** Let  $k(x) = (1 - x^2) \cdot \mathbb{I}[x \in [-1, 1]] = \begin{cases} 1 - x^2 & x \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$  be the kernel of a continuous distribution.

- a) Determine the pdf for this distribution.
- b) Compute the mean and variance for this distribution

**2.4** Let  $Y \sim \text{Triangle}(0, 10, 4)$ . Compute  $E(Y)$  and the median of  $Y$ .

**2.5** Let  $W \sim \text{Unif}(0, 10)$ . Compute  $E(W)$  and  $\text{Var}(W)$ .

### 2.6

- a) Let  $X \sim \text{Exp}(4)$ . Use R to compute  $E(X)$ .
- b) Let  $X \sim \text{Exp}(10)$ . Use R to compute  $E(X)$ .
- c) Let  $X \sim \text{Exp}(1/5)$ . Use R to compute  $E(X)$ .
- d) What pattern do you notice. Explain in terms of the definition of the exponential distribution why this makes sense.

**2.7** Use R to plot the pdf and compute the mean and variance of each of the following distributions.

- a) `Beta(2, 3)`
- b) `Beta(20, 30)`
- c) `Gamma(shape = 2, scale = 3)`

d) `Weibull(shape = 2, scale = 3)`

**2.8** For each of the following distributions, determine the proportion of the distribution that lies between 0.5 and 1.

a) `Exp(rate = 2)`

b) `Beta(shape1 = 3, shape2 = 2)`

c) `Norm(mean = 1, sd=2)`

d) `Weibull(shape = 2, scale=1/2)`

e) `Gamma(shape = 2, scale=1/2)`

## 2.9

- a) Using Table 2.1 and the method of moments, fit an exponential distribution to the Twin Falls wind speed data. What is the estimated value of the rate parameter?
- b) Now use `fitdistr()` to fit an exponential distribution using maximum likelihood.
- c) How do the two estimates for the rate parameter compare?
- d) How well does an exponential distribution fit this data?

**2.10** A Gamma distribution can also be fit using the method of moments. Because there are two parameters (shape and rate or shape and scale), you will need to solve a system of two equations with two unknowns.

- a) Using Table 2.1 and the method of moments, fit a Gamma distribution to the Twin Falls wind speed data. What are the estimated values of the shape and rate parameters?
- b) How do the method of moments estimates for the parameters compare to the maximum likelihood estimates from `fitdistr()`?

**2.11** Sam has found some information about wind speed at a location he is interested in online. Unfortunately, the web site only provides the mean and standard deviation of wind speed.

mean: 10.2 mph  
standard deviation: 5.1 mph

- a) Use this information and the method of moments to estimate the shape and rate parameters of a Gamma distribution.
- b) In principal, we could do the same for a Weibull distribution, but the formulas aren't as easy to work with. Fit a Rayleigh distribution instead (i.e., a Weibull distribution with shape parameter equal to 2).

**2.12** In 1964, a study was undertaken to see if IQ at 3 years of age is associated with amount of crying at newborn age. In the study, 38 newborns were made to cry after being tapped on the foot, and the number of distinct cry vocalizations within 20 seconds was counted. The subjects were followed up at 3 years of age and their IQs were measured. You can load this data using

```
Baby <- read.file("http://www.calvin.edu/~rpruim/data/BabyCryIQ.csv")

## Reading data with read.csv()

head(Baby)

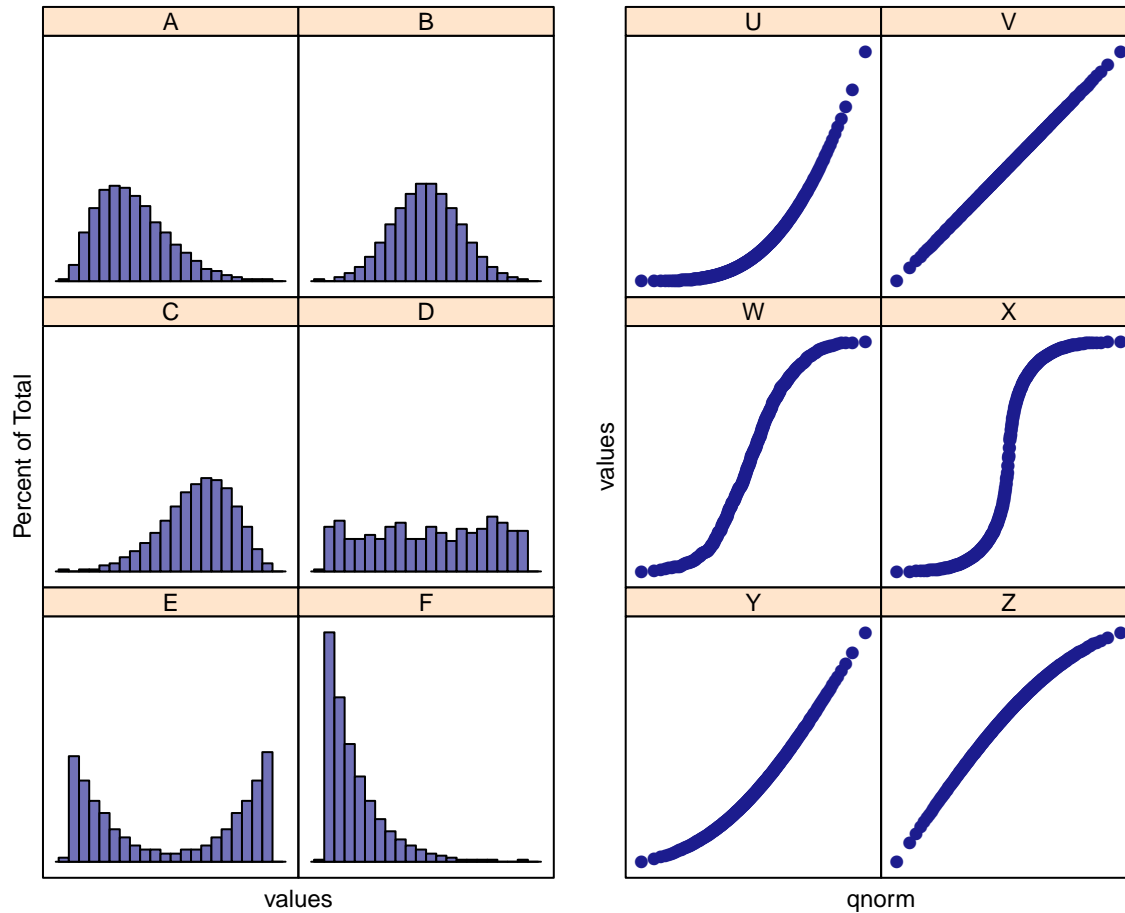
##   cry.count  IQ
## 1      10  87
## 2      20  90
## 3      17  94
## 4      12  94
## 5      12  97
## 6      15 100
```

The `cry.count` variable records the number of distinct cry vocalizations within 20 seconds. Choose a family of distributions to fit to this data and do the fit using `fitdistr()`. Also include a plot showing a histogram and your fitted density curve.

**2.13** Create normal quantile plots for the ages of patients in the `HELPrct` data set separated by `substance`. (Getting separate or overlaid plots using `qqmath()` works just like it does for other lattice plots).

Comment on the plots.

**2.14** Match the normal-quantile plots to the histograms.



**2.15** Show that  $\text{Var}(X) = E(X^2) - E(X)^2$  by showing that

$$\int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu_X^2$$

whenever  $f$  is a pdf and all the integrals involved converge.

**2.16** The heights of 18–22 year olds in the US follow approximately normal distributions within each sex. Estimated means and standard deviations appear in the table below.

	mean	standard deviation
women	64.3 in	2.6 in
men	70 in	2.8 in

Answer the following questions without using a computer or calculator (except for basic arithmetic).



- a) If a woman is 68 inches tall, what is her z-score?
- b) If a man is 74 inches tall, what is his z-score?
- c) What is more unusual, a woman who is at least 68 inches tall or a man who is at least 74 inches tall?
- d) Big Joe has decided to open a club for tall people. To join his club, you must be in the tallest 2.5% of people of your sex. How tall must a woman be to join Big Joe's club?
- e) How tall must a man be to join Big Joe's club?

**2.17** Use the information from the previous problem to answer the following questions.

- a) What proportion of women are 5'10" or taller?
- b) What proportion of men are 6'4" or taller?
- c) If a man is in the 75th percentile for height, how tall is he?
- d) If a woman is in the 30th percentile for height, how tall is she?

