# Chapter 1 Random Variables, Sampling and Distributions

## 1.1 Learning Objectives for Chapter

1. Define random variables and their distributions.
2. Define and identify population, sample and observation.
3. Explain why we need to sample populations.
4. Describe and use a method to obtain random samples with equal probabilities.
5. Define “parameter” and “random variable” and discuss their relationship to one another
6. Describe and implement stratified random sampling.
7. Explain the difference between sampling with and without replacement.
8. Write and use equations for estimators of population mean and variance.
9. Distinguish between estimators and definitions of parameters.
10. List the most important distributions in elementary statistics.
11. Define and calculate bias of an estimator of a parameter.
12. When should a Student’s t distribution be used over a normal distribution for samples?
13. Define statistical bias and give a few examples of situations that may lead to a biased estimator
14. Sketch an example sampling distribution (r = 10) for observations drawn from a normally distributed population, and sketch the corresponding normal pdf, using your sampling distribution to estimate its parameters.

- List the differences and similarities between the two sketches.
- Explain why the normal pdf representing the population from which the samples were drawn might not look like the normal pdf you drew.

## 1.2 Random variables

Random variables are those variables whose values are determined by random processes or experiments. Any time we conduct a measurement or observe the result of an experiment we obtain specific values for the random variables. These specific values are termed realizations. For example, consider the result of rolling a pair of dice. Define the random variable as the total number of dots in both dice. The random variable can take integer values between 2 and 12. Say that in one roll you got a 3 and a 5. The realization of the random variable is 8.

In reality, a random variable is a real-valued function defined over the sample space of a random experiment or process. We are free to define the function in any way we want, even with a table that shows the value for each outcome, but because it has to be a function, each result of the random process cannot be associated with more than a single value of each random variable.

- A random variable is a function that associates a single real number with each outcome in the sample space of a random experiment.

Do not confuse this function property with the posibility of having random variables grouped into vectors, where each vector contains more than one value, each value being the realization of a different random variable. For example, in the roll of two die we can define the random vector {number of dots in the first die, number of dots in the second die}. The importance of thinking of a random variable as a function is that it give a lot of flexibility to deal with the results of random experiments. Instead of just counting dots on the face of a die we can use complicated functions that allow us to test hypotheses about more complex processes, such as administering medicines to animals, or comparing the yield of multiple varieties of food plants.

A more formal definition and an intuitive physical model of random variable is given in [@286171]. Make sure to also explore the explanations given by W. Huber in [@54894].

**Examples of Random Variables**

- Number of dots in three dice rolled once.
- Number of plants within 50 cm of a point placed randomly in a wheat field.
- Time that it takes to find a bird of a certain species in a forest plot.
- Your diastolic blood pressure measured every day at noon.
- Weight of a randomly selected mature dairy cow in California.
- Number of atoms that decay in 1 second in 1 mg of Uranium-238.
- Pain level reported by a patient who consults the doctor due to a headache.
- Weight of a ball randomly selected from two.
- Mass of carbon in a soil sample divided by the total sample dry mass.
- Natural log of nonstructural carbohydrates concentration in CA almond tree branches on 1 Mar every year.

### 1.2.1 Types of Variables and Notation

The choice of statistical method to answer a specific question depends strongly on the types of random variables involved in the question. The type of variable involved matters primarily because it determines the type of statistical distribution they can have.

There are several ways to classify variables, but the following is perhaps the most useful and general. Statistical variables can be:

**Categorical** For example, plant species or cultivar. These are non-numeric variables whose values are not ordered. The order of values has no meaning.

**Ordinal** For example, plant growth stages or disease condition. Ordinal variables are non-numeric but their values are ordered and the order has meaning. We know that an animal in stable disease condition is worse than healthy and better than gravelly ill, but it is not possible to say if “healthy” is 10 or 1.5 times better than gravelly ill.

**Quantitative** These are random variables that take numerical values and that can be used in arithmetic operations such as addition and multiplication. Quantitative variables can be discrete or continuous.

**Discrete** Discrete quantitative variables can take integer values, for example, the number of branches in a plant, or the number of eggs that a hen produces. The values come from the set of Integer Numbers. Discrete random variables can have finite or infinite sampling spaces. For example, the number of heads in 37 coin tosses is discrete and has a finite sampling space $S = \{0, 1, \ldots, 37\}$, whereas the number of tosses until 3 heads are in a row is discrete but has a countable infinite sampling space $S = \{3, 4, \ldots \infty\}$.

**Continuous** Continuous quantitative variables can take values between any other two values. For example, the mass of wheat produced per unit area, or the volume of milk produce by a cow per day. These variables take values that are Real Numbers. Note that most of the specific variables we analyze in agricultural and environmental sciences are positive real numbers.

### 1.2.2 Using random variables

Once we define a useful random variable, we can apply the rules of probability to calculate the probability of any event in terms of the random variable, but first we have to know the statisticsl distribution of the random variable. Imagine that you are breeding animals and need to get at least three individuals with a specific genotype of class G that is expected to appear in 20% of the offspring. What are the chances (probability) that you will get the 3 G's if you get 5 offspring? We can *model* this situation assuming that the genotypes of all individuals are independent, and that the probability for each individual having G is 0.20. Let's use the letter "O" for "other" to designate individuals that are not G. You will have the 3 you need if you get 3, 4 or all 5 with G. We define the random variable Y as "number of G's in five offsping." If you get 3 you do not get 4 or five. This means that the events are disjoint and that we sum their probabilities to get the probability of at least 3 G's. To calculate the probability of each of the events, we recall that individuals are independent, so the probability of a given set, like GGGOO is the product of the individual probabilities, in this case $0.20^3 \cdot 0.80^2$. Finally, we take into account that the event GGGOO is not the same as GGOOG, but each has probability $0.20^3 \cdot 0.80^2$. The ennumeration of the combinations and probabilities show how a pattern emerges:

Table: (#tab:deriveBinomialExmpl) Calculation of the probability of obtaining Y successes in 5 independent trials with constant probability of success.

| Y | sequence | no. sequences | Probability of each sequence |
|---|---|---|---|
| 0 | OOOOO | 1 = $C_0^5$ | $0.20^0 \cdot 0.80^5 = 0.32768$ |
| 1 | GOOOO, OGOOO, OOGOO, OOOGO, OOOOG | 5 = $C_1^5$ | $0.20^1 \cdot 0.80^4 = 0.08192$ |
| 2 | GGOOO, GOGOO, GOOGO, GOOOG, OGGOO, OGOGO, OGOOG, OOGGO, OOGOG, OOOGG | 10 = $C_2^5$ | $0.20^2 \cdot 0.80^3 = 0.02048$ |
| 3 | GGGOO, GGOGO, GGOOG, GOGGO, GOGOG, GOOGG, OGGGO, OGGOG, OGOGG, OOGGG | 10 = $C_3^5$ | $0.20^3 \cdot 0.80^2 = 0.00512$ |
| 4 | GGGGO, GGGOG, GGOGG, GOGGG, OGGGG | 5 = $C_4^5$ | $0.20^4 \cdot 0.80^1 = 0.00128$ |
| 5 | GGGGG | 1 = $C_5^5$ | $0.20^5 \cdot 0.80^0 = 0.00032$ |

There are many ways to get any number, say r, G's in a sequence. That number equals the number of different sets of r positions or slots that can be chosen from the 5 slots available, without replacement. This is a case where the order in which the slots are chosen does not matter (having G's in positions 1 and 3 is the same as having G's in postions 3 and 1) and there is no replacement (one a slot is occupied it cannot be selected again). Therefore, we can use combinations of 5 slots taken in sets of r = Y to see to determine the number of ways that a given number of successes can be obtained. Notice that whenever you get r slots with G's you also 5-r slots for O's. That's why the number of sequences is symmetric from top to bottom of the table.

If we use the letters $p$ for the probability of G or "success," $q = 1 - p$ for failure or O and $n$ for the total number of trials we can write a compact equation for the probability that we obtain Y = r successes. We use the letter $Y$ for the random variable "number of successes" to emphasize that it is a response variable, but we also use the letter $r$ to link this to Equation (??) for combinations and to designate a specific realization of Y.

$$P(Y = r) = C_r^n \; p^r \; q^{n-r} \tag{1.1}$$

This equation is the function that associates a probability to each value of the random variable and it is the *probability mass function* for the **Binomial Distribution**. The term "binomial" comes from the fact that the equation represents all the terms in the expansion of the power of a sum of the form $(p + q)^n$. The binomial expansion states that $(p + q)^n = \sum_{r=0}^{n} \; p^r \; q^{n-r}$. This equation can be used to answer any questions that involve Binomial distributions. Specifically, the probability of getting at least r G's in the example is $P(Y \geq 3) = \sum_{r=3}^{n} \; 0.20^r \; 0.80^{5-r}$

**Functions of random variables**

Suppose that in the example above, animals with G require 2 tons of feed to reach maturity, whereas animals O require 1.5 tons. How much feed do you need to order to have a probability of 90 or better that you will have enough to feed all animals to maturity? We can define a new random variable based on the first one:

$$X = 2\,r + 1.5\,(5 - r) = 7.5 + 0.5\,r$$

```
an.dat <- data.frame(Y = 0:5) # list of possible values of rv Y

an.dat$Prob <- dbinom(an.dat$Y, size = 5, prob = 0.20) # corresponding porbabi

an.dat$X = 7.5 - 0.5 * an.dat$Y # lisf of possible values of rv X

an.dat$cdf <- cumsum(an.dat$Prob) # cumulative probabilities

kable(an.dat, caption = "Probability mass function and cumulative distribution
   kable_styling(full_width = FALSE)
```

Table 1.2: Probability mass function and cumulative distribution function (CDF) for a binomial random variable Y with p = 0.20 and n = 5. Random variable X is a linear function of Y

| Y | Prob | X | cdf |
|---|---:|---:|---:|
| 0 | 0.32768 | 7.5 | 0.32768 |
| 1 | 0.40960 | 7.0 | 0.73728 |
| 2 | 0.20480 | 6.5 | 0.94208 |
| 3 | 0.05120 | 6.0 | 0.99328 |
| 4 | 0.00640 | 5.5 | 0.99968 |
| 5 | 0.00032 | 5.0 | 1.00000 |

Using the relationship between X and Y we obtain the probability mass function for X, which is just the same as the original but translated (by adding 7.5) and shrunk (by multiplying by 0.5). By progressively adding the probability of each value of X, starting from the lowest one we obtain the cumulative density function, which gives us the probability that X is equal to or less than r: $F(r) = P(X \leq r)$. We can see that in order to have a probability greater than or equal to 0.90 of having enough feed we need to order 8.5 tons of feed (Table 1.2).

## 1.3 Probability Distributions

Probability distributions are functions of random variables that have the following properties:

**Discrete Random Variables** have probability mass functions $P(Y = y)$

$$\begin{aligned} P(Y = y) & \quad \text{is the probability of the event that results in } Y = y \\ P(Y = y) & \geq 0 \\ \sum_{y=-\infty}^{y=+\infty} P(Y = y) & = 1 \end{aligned} \tag{1.2}$$

The corresponding **cumulative distribution function** is

$$F_Y(y) = P(Y \leq y) = \sum_{i=-\infty}^{i=y} P(Y = i) \tag{1.3}$$

**Continuous Random Variables** have probability density functions $f(Y = y)$

$$f(y) \quad \text{is the probability "per unit change of Y"}$$

$$P(Y = y) = 0 \quad \text{the probability of any specific value y is zero!}$$

$$P(a \leq Y \leq b) = \int_a^b f_Y(y)\, dy$$

$$\int_{\mathbb{R}} f_Y(y)\, dy = 1 \tag{1.4}$$

The corresponding **cumulative distribution function** is

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f_Y(y)\, dy \tag{1.5}$$

The symbol $\int_{\mathbb{R}}$ represents the *integral* of the function over all real values of Y, and $\int_{\mathbb{R}} f(y)\, dy$ is the integral between $Y = a$ and $Y = b$. If you are not familiar with the concept of integral, no worries. Imagine that $f(y)$ is an abolutely continuous function of the value $y$. The integral between any two $y$ values a and b is the area under the curve between the two points (Figure 1.1). One way to calculate the area approximately is to divide the interval [a, b] into $(b - a)/dy$ little segments. The area of each segment is the base $dy$ times the height, which can be set at $f(y_{lo} + dy/2)$ where $y_{lo}$ is the left end of each segment.¬

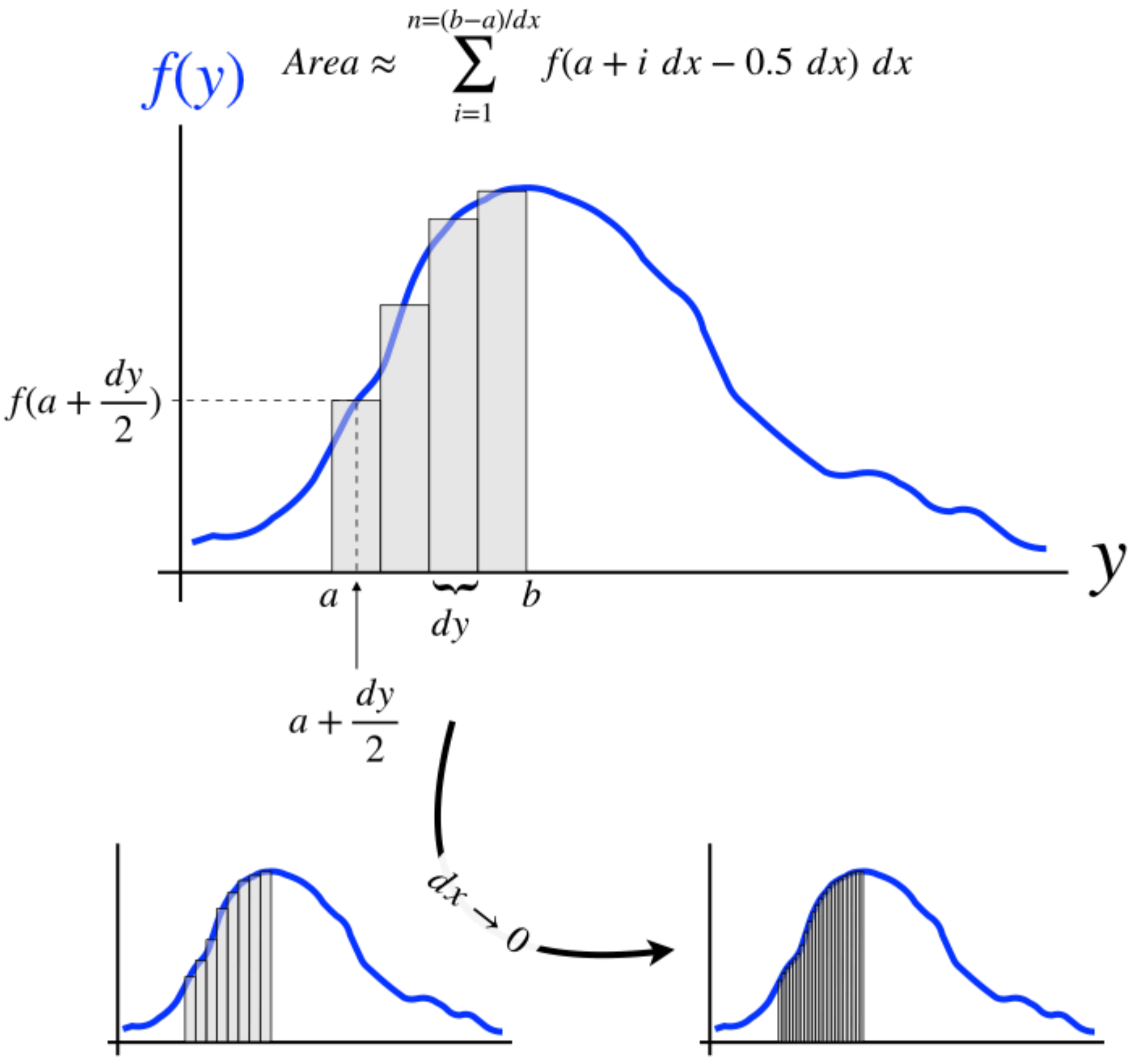


Figure 1.1: The integral of a continuous probability density function (pdf) can be though of as the limit of the sum of rectangles with base width $dy$ and height equal to the height of the pdf function at the center of the base, as $dx$ approaches zero. The area of each rectangle is an approximation of the probability that Y takes values within the base of the rectangle.

- Probability distributions are functions that define the probability of any event occurring in a random experiment where the outcome is the value of a random variable.
- For example, if the random experiment is flipping a fair coin the probability of X = 1 (heads) is the same as probability of X= 0 (tails) and it is 0.50. Then, X is a random variable that has a Bernoulli distribution with parameter p = 0.5.
- Once we know the true distribution of a random variable, we can know everything there is to know about the random variable.

- True distributions for real random variables are rarely known. Even if we assume that we know the distribution family we never know the parameter values.
- In the case of the coin, we know the distribution of a theoretical coin that is fair, perfect and eternal. We use it as a model for real coins, but we know it is not a perfect model. Some people are able to flip any number of heads or tails they want. We can also suspect that each type, or even each individual coin, has a probability of heads that is different and not exactly 0.50. The same can be said of dice. Recall that dice can be loaded and/or imperfect. However, the usual distributions used for them are *very good models* for the corresponding random experiments and variables.

## 1.4 Parameters and Moments

Distribution functions are usually functions that can be written down with equations that have certain constant values for all values of the random variable. The set of all possible values of the random variable is called the *support*. In the example about breeding animals above, Equation (1.1) has two parameters: n, the number of trials and p, the probability of success; q does not count as a different parameter because it is simply $1 - p$. All distributions will be characterized by their parameters and other properties such as the mean or expected value and variance. The mean, variance, skeweness and kurtosis are called ther **moments** of a distribution, because they all arise from variations of the same equation.

This section presents the definitions of the expectation, variance and parameters of distributions using equations that may be a little scary at first. Students are not expected to be able to derive these equations, and in many cases not even to memorize them. The main point in presenting them is for students to realize and remember that moments and parameters are characteristics of the whole distribution, or of populations, whereas averages and sample variances are calculated from samples and use **different equations**. Moreover, for any given random variable, the parameters and moments of its distribution are *constants*, whereas averages and variances estimated with samples from the distribution are themselves random variables that change from sample to sample. This is a very important concept.

### 1.4.1 Expectation, Mean or First Moment

The expected value of a random variable is the long-run average of values resulting from repeating the random experiment a very large number of times. According to the **Law of Large Numbers**, the average of a sequence of independent draws from the same random variable will tend towards the mean of the random variable as the number of draws increases. The expectation, represented by $E\{\}$ or mean is more formally defined as follows:

**Expectation of discrete random variables**

$$E\{Y\} = \sum_i y_i \, P(Y = y_i) \tag{1.6}$$

The expectation of the number of animals with genotype class G above is:

$$\begin{aligned} E\{Y\} = \mu_Y &= 0 \times 0.32768 \\ &+ \ 1 \times 0.40960 \\ &+ \ 2 \times 0.20480 \\ &+ \ 3 \times 0.05120 \\ &+ \ 4 \times 0.00640 \\ &+ \ 5 \times 0.00032 = 1 \end{aligned}$$

**Expectation of continuous random variable**

$$E\{Y\} = \int_{\mathbb{R}} y \, f_Y(y) \, dy \tag{1.7}$$

The integral appears again. In this case, think of it as weighted average of values of Y at the center of a large number of small intervals that cover the whole horizontal axis. The weigthing factor is the area under the curve inside each narrow column, which is approximated by the product of the width of the interval and height if the curve at the center. As you make the intervals $dy$ smaller and smaller, the calculation approaches the integral.

As an example for the calculation of the mean of a continuous variable, let’s use the continuous uniform distribution between real numbers a and b. The probability density function of this function is rather simple:

$$f(y) = \begin{cases} \frac{1}{b-a} & \text{if } a \le y \le b \\ 0 & \text{otherwise} \end{cases} \tag{1.8}$$

This pdf is zero outside the interval $[a, b]$ and constant in the interval Because the area has to be 1.0, then, the height of the rectangle and the value of the function are constant at $1/(b-a)$. Using a bit of calculus or intuition we obtain the expectation:

$$E\{Y\} = \int_a^b \frac{y}{b-a}\,dy = \frac{1}{b-a}\int_a^b y\,dy = \frac{1}{b-a}\,\frac{y^2}{2}\bigg|_a^b = \frac{b^2-a^2}{2\,(b-a)} = \frac{a+b}{2}$$

Thus, the expectation or mean of a random variable with a continuous uniform distribution between 3 and 7 is 5, the midpoint of the distribution.

### 1.4.2 Variance or Second Moment

The variance of a distribution, represented by $V\{Y\} = \sigma_Y^2$ can be defined as the first moment or expectation of the squared deviations from the mean:

$$V\{Y\} = E\{(Y - E\{Y\})^2\} = E\{(Y - \mu)^2\} \tag{1.9}$$

This definition is short and easy to remember, and it will also remind us of the equation to estimate the variance based on a sample, where we divide the sum of the squared deviation by the number of observations.

- The variance is the expectation of squared deviations from the mean.
- The standard deviation is the square root of the variance.
- Both are always positive.

An important property of the expectation is that the expected value of a function $g(Y)$ of a random variable can be obtained simply by using the definition of expectation where instead of Y we plug in $g(Y)$. This is called the *Law of the Unconscious Statistician* or LOTUS. The application of Equation (1.9) and the LOTUS, with $g(Y) = (Y - \mu)^2$ to discrete distributions yields the definition of variance for discrete distributions:

$$V\{Y\} = E\{(Y - E\{Y\})^2\} = E\{(Y - \mu)^2\} = \sum_i (y_i - \mu)^2 \ P(Y = y_i) \qquad (1.10)$$

In the case of continuous distributions, the variance becomes

$$V\{Y\} = E\{(Y - \mu)^2\} = \int_{\mathbb{R}} (y - \mu)^2 \ f_Y(y) \qquad (1.11)$$

### 1.4.3 Covariance and Correlation between two RV’s

Recall the definitions of marginal, joint and conditional probability from the section on Probability of two events. Consider two random variables together, and think of the values they take as the events. We can then calculate the joint and conditional probabilities for any combination of values from the two variables. This is easily illustrated with discrete finite random variables, but the concepts extend to any pair of random variables.

Imagine that we inspect the tidal zone in a segment of coast (your population) and count the number of starfish (X) and large sea snails (Y) in each 2 x 2 $m^2$ of the tidal zone during low tide. Because we define the area at the time of measurement as our whole population, we know the complete population and we can calculate the moments. The table of (fictitious) results expressed as joint relative frequencies is:

Table 1.3: Joint and marginal probabilities of number of starfish (columns) and number of sea snails (rows) in each 2 x 2 m-sq of a tidal zone. (Fictitious data)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | Sum |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.1363 | 0.0266 | 0.0033 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.1662 |
| 1 | 0.1360 | 0.1311 | 0.0250 | 0.0020 | 0.0002 | 0.0000 | 0.0000 | 0.2943 |
| 2 | 0.0668 | 0.1226 | 0.0678 | 0.0116 | 0.0009 | 0.0002 | 0.0000 | 0.2698 |
| 3 | 0.0217 | 0.0569 | 0.0560 | 0.0215 | 0.0038 | 0.0000 | 0.0000 | 0.1600 |
| 4 | 0.0061 | 0.0184 | 0.0242 | 0.0166 | 0.0039 | 0.0016 | 0.0000 | 0.0709 |
| 5 | 0.0014 | 0.0053 | 0.0090 | 0.0073 | 0.0034 | 0.0011 | 0.0001 | 0.0276 |
| 6 | 0.0001 | 0.0009 | 0.0025 | 0.0020 | 0.0021 | 0.0006 | 0.0003 | 0.0085 |
| 7 | 0.0001 | 0.0002 | 0.0005 | 0.0007 | 0.0006 | 0.0003 | 0.0002 | 0.0026 |
| Sum | 0.3685 | 0.3621 | 0.1884 | 0.0617 | 0.0149 | 0.0038 | 0.0006 | 1.0000 |

The mean or expectation for number of starfish (X) is the sum of each number of starfish multiplied by its marginal frequency:

$$\begin{aligned} E\{X\} = \sum_{x=1}^{6} x\, P(X = x) = 0 \times 0.3685 \\ + 1 \times 0.3621 \\ + 2 \times 0.1884 \\ + 3 \times 0.0617 \\ + 4 \times 0.0149 \\ + 5 \times 0.0038 \\ + 6 \times 0.0006 = 1.0063 \end{aligned}$$

As an exercise, fully expand the calculation for the snails mean, ($E\{Y\} = 1.8048438751$).

Inspection of Table 1.3 reveals that the events are not independent. The joint probabilities are not the product of the marginal probabilities. We can get the conditional probabilities for each number of snails given the number of starfish by dividing each column by the marginal column sum at the bottom of Table 1.3. The results are displayed in a graph so the differences can be seen at once.

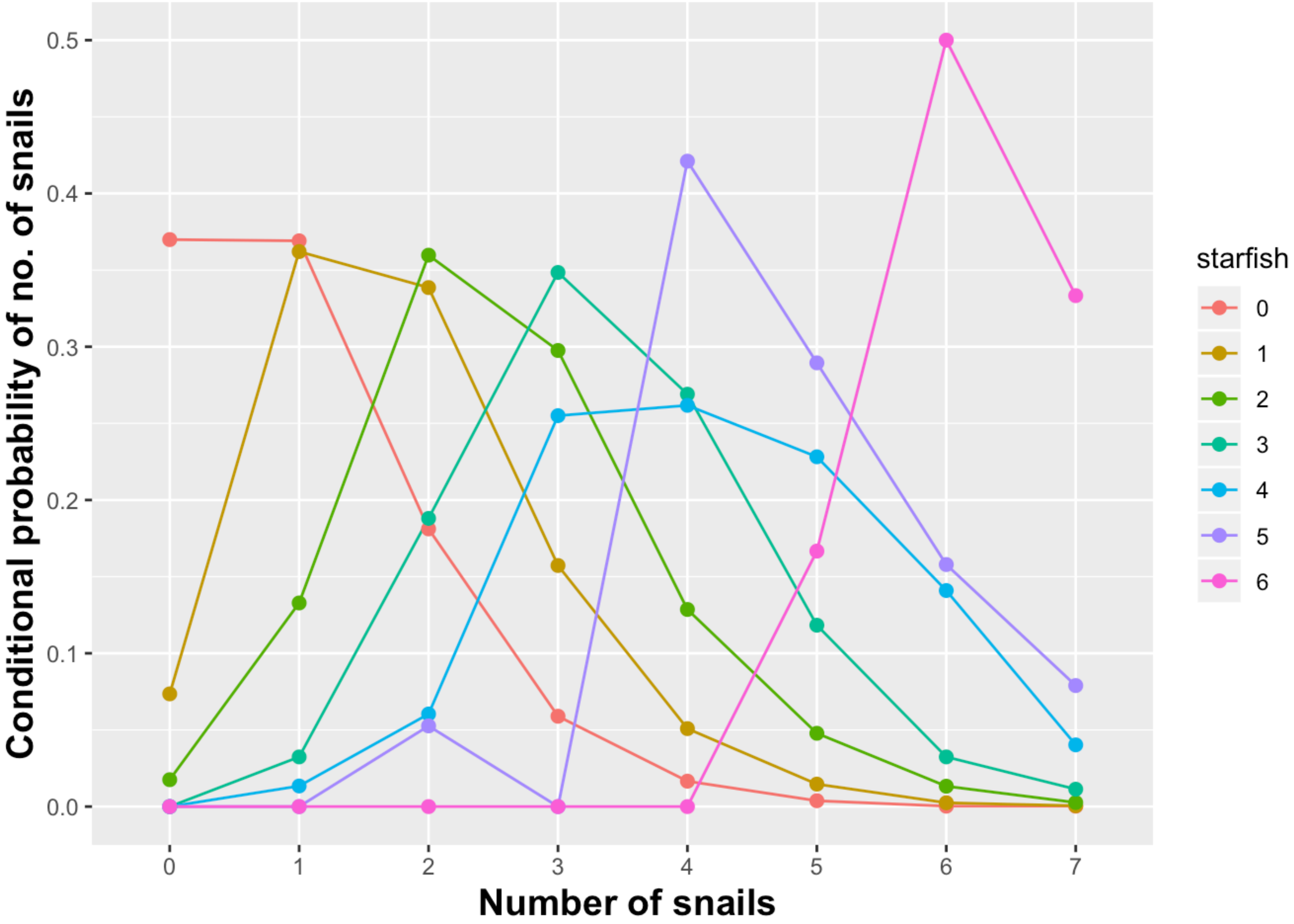


Figure 1.2: Conditional probability of number of snails, given a number of starfish. The distributions are discrete and have no height for axis values between integers. Lines are added just to facilitate the indentification of the different sets of points.

As the number of starfish increases, the distribution of the number of snails moves to the right, therefore, the two random variables are not independent. The level of dependence between two random variables is measured by their **covariance**. The covariance between two random variables is defined as the expectation of the cross product of deviations of each variable from its mean.

**Covariance**

- The covariance between two random variables is the expectation of the products of deviations from each observation to the mean in each variable.

$$
\begin{aligned}
V\{X,Y\} = \sigma\{X,Y\} = cov\{X,Y\} &= E\{(X - E\{X\}) \cdot (Y - E\{Y\})\} \\
&= E\{(X - \mu_X) \cdot (Y - \mu_Y)\} \qquad (1.12)
\end{aligned}
$$

In the case of discrete random variables like the number of starfish and snails, we apply the definition of expectation for discrete variables to obtain:

$$V\{X,Y\} = \sum_i \sum_j (x_i - \mu_x)(y_j - \mu_y)\; P(X = x_i, Y = y_j)$$

Plugging in some of the values from the example we get

$$
\begin{aligned}
V\{X,Y\} = &(0-1.006)\times(0-1.805)\times 0.1363 \\
&+(1-1.006)\times(0-1.805)\times 0.0266 \\
&+(2-1.006)\times(0-1.805)\times 0.0033 \\
&\ldots \\
&+(0-1.006)\times(1-1.805)\times 0.1360 \\
&+(1-1.006)\times(1-1.805)\times 0.1311 \\
&+(2-1.006)\times(1-1.805)\times 0.0250 \\
&\ldots \\
&\ldots \\
&+(4-1.006)\times(7-1.805)\times 0.0006 \\
&+(5-1.006)\times(7-1.805)\times 0.0003 \\
&+(6-1.006)\times(7-1.805)\times 0.0002 = 0.8088
\end{aligned}
$$

As an exercise, complete the calculation using R and corroborate that the result 0.8088 is correct. The covariance is positive, indicating that both variables tend to increase or decrease together. This can be seen in a graph showing the **bivariate** distribution of number of starfish and snails.

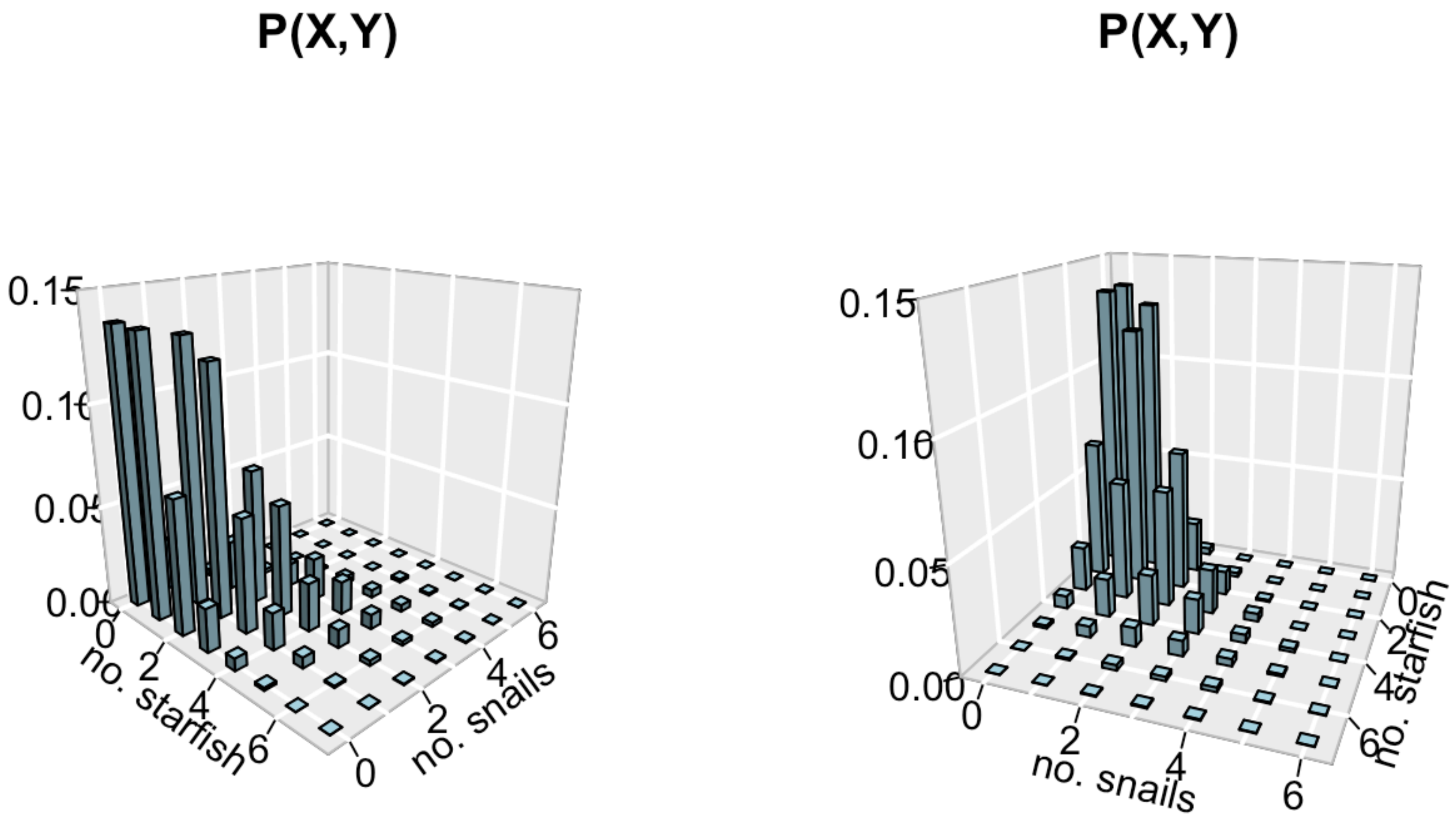


Figure 1.3: Two perspectives of the bivariate distribution of two random variables, number of starfish and number of snails in all quadrats of a tidal zone. Although bars are used to represent the joint probabilities, in reality the bars should only have height, with null width and lenght, because both random variables are discrete.

When variables are continuous, the application of the expectation operator to the definition of covariance yields equations with double integrals. We extend the idea of integral in one dimension, explained in Figure 1.1, to two dimensions. The probability density function is now a surface in two dimensions and the volume under it is divided into columns whose widths tend to zero. The probability of any region of X and Y values is the volume under the surface inside the area defined by the region.

$$V\{X,Y\} = \int_x \int_y (x - \mu_x)\,(y - \mu_y)\,f(x,y)\,dx$$

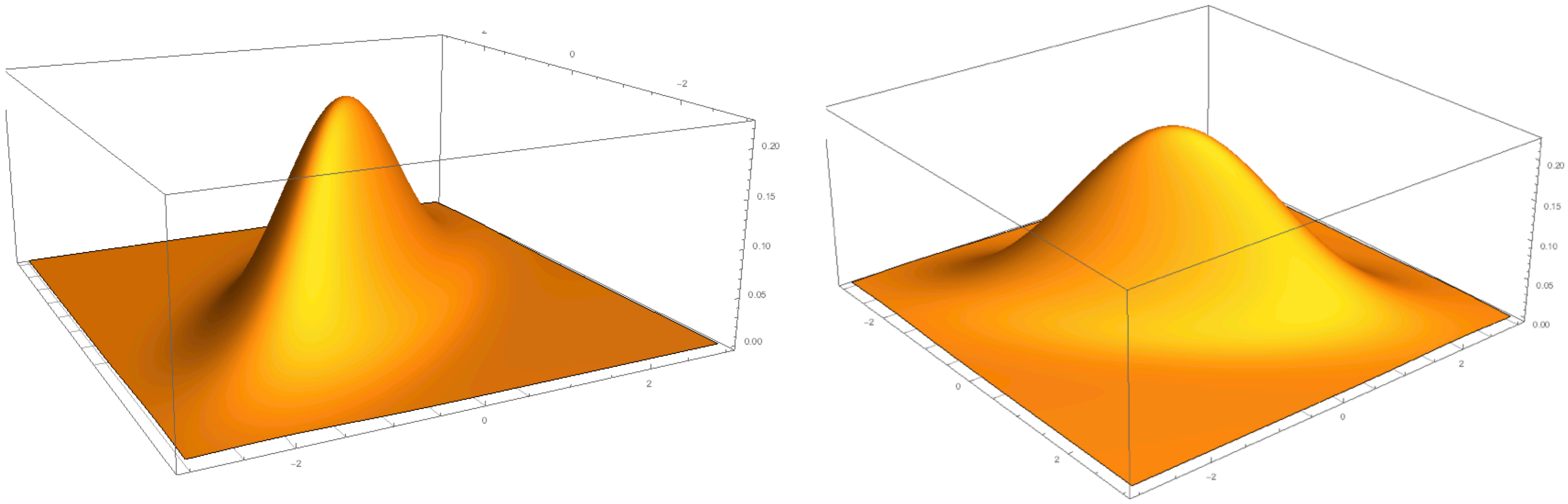


Figure 1.4: Two perspectives of the joint probability density function for two normal random variables that have a correlation of 0.70. The height of the surface is the probability per unit volume under the surface. The intersection between any vertical plane and the surface is the pdf of a normal distribution. In particular,intersections with planes perpendicular to either one of the axes (variables) are the conditional distributions of the other variable.

### Correlation

The covariance has a drawback to representt the degree of association between two random variables because it is a quantity with units and its numerical value depends on the magnitude of the random variables and the units. Therefore, covariances cannot be compared when variables have different units or when they have different means. The **correlation** coefficient is a better measure of association between two variables because it has no units and can be compared across variables and situations.

- The correlation between two random variables equals their covariance divided by the product of their standard deviations.

$$cor\{X,Y\} = r_{X,Y} = \frac{V\{X,Y\}}{\sqrt{V\{X\}\,V\{Y\}}} \tag{1.13}$$

When random variables are standardized (see below), their covariance and correlation are equal.

- When random variables are independent, their covariance and correlation are zero, but the converse is not true. Random variables with zero correlation may have nonlinear dependence.

### 1.4.4 Parameters are not necessarily the moments

Distribution functions have constant values that are part of the function called **parameters**. In order to know everything about any given distribution, one must know all the parameters. The Uniform distribution seen before has two parameters, a and b. The Binomial distribution also has two paramters, n and p. As we will see in more detail below, the Poisson distribution has a single parameter. In the case of the Uniform and Binomial distributions, their means and variances are not the same as the parameters. In the case of the Poisson distribution, the mean and the variance both are equal to the single parameter. Other distributions, like the Gamma distribution have more than two parameters. The Normal distribution has two parameters, its mean and its variance; it is a very special case.

The point is that parameters are not necessarily the same as mean and variance. The Normal distribution is a special case in which there are only two parameters and they happen to be the mean and the variance. Therefore, in order to know everything about a Normal distribution one only needs two numbers: mean and variance.

### 1.4.5 Properties of Mean and Variance

The following properties can be easily derived directly from the definitions of mean and variance, and they are very useful when working with samples. Consider that $X$, $Y$ and $Z$ are random variable with any distributions that has mean and variance[1].

$$
\begin{aligned}
&E\{a + b\,Y\} = a + b\,E\{Y\} \\
&V\{a + b\,Y\} = b^2\,V\{Y\} \\
&\text{Let X be a linear combination of random variables Y and Z:} \\
&X = a + b\,Y + c\,Z \quad \text{then,} \\
&E\{X\} = a + b\,E\{Y\} + c\,E\{Z\} \\
&V\{X\} = b^2\,V\{Y\} + c^2\,V\{Z\} + 2\,b\,c\,V\{Y, Z\} \\
&\text{Special cases for independent variables:} \\
&V\{Y + Z\} = V\{Y - Z\} = V\{Y\} + V\{Z\}
\end{aligned}
\tag{1.14}
$$

The last equation is extremely important in statistics. In most real applications of statistics we have to use models that involve estimation of more parameters than just mean and variance. For example, in linear regression we will estimate intercept and slope. **Estimated parameters** are based on samples of random variables, so they are themselves random variables. When we use the estimated intercept and slope to make a prediction, we are using a linear combination of estimated parameters. The last line in Equation (1.14) allows us to estimate the variance of the prediction and thus make confidence intervals or test hypothesese that involve linear combinations of estimated parameters.

- A main goal of statistics is to make statements about linear combinations of estimated parameters.
- The variance of a linear combination of estimated parameters depends on the variances and covariances of the estimated parameters.

### 1.4.6 Standardized variables

Random variables are frequently standardized to obtain new random variables with more desirable properties. Standardization is a linear transformation or function of the random variable by which the mean is subtracted and the result is divided by the standard deviation.

Frequently we use the lower case version of a letter to refer to the standardized variable. The letter “z” is typically used to refer to the standard Normal distribution.

The mean and variance of a standardized variable are obtained directly by the properties of the mean and variance treated in the previous section. Standardized random variables have mean zero and variance one.

$$\begin{gathered}
X \; \text{ is an RV with } \; E\{X\} = \mu_X \text{ and } V\{X\} = \sigma_X^2 \\
\Longrightarrow \; x = \frac{X - \mu_X}{\sigma_X} = -\left(\frac{\mu_X}{\sigma_X}\right) + \left(\frac{1}{\sigma_X}\right) \; X \\
\text{is an RV with } \; E\{X\} = 0 \text{ and } V\{X\} = 1
\end{gathered} \tag{1.15}$$

The function of X used to get x is a *linear* function, where $\mu_X$ and $\sigma_X$ are constants that do not depend on X. Linear functions simply translate (move along the axis) and uniformly stretch or shrink the width and height of the distribution function.

## 1.5 Common distributions

The list of distribution functions is infinite. In theory, random variables can have any distribution, for as long as they comply with the definition, which is rather general. In practice, we manage to make useful models of the world with very few prototypical families of distributions. Loosely, a family of distributions is a set of distributions that differ only in the values of their parameters, but have the same functional form. For example, the Bernoulli distribution, which has value 1 with probability $p$ and 0 with probability $1-p$, is a Binomial distribution where $n=1$.

Some distributions can be derived on the basis of a few facts or assumptions. For example, the Binomial distribution results from performing a set number of independent trials with constant probability of success. Other distributions may be the result of complicated combinations of basic ones.

In this section we describe the distributions that we will use in the rest of the book. Wikipedia has entries for about 30 discrete and 100 continuous distributions, where they are nicely organized and described.

### 1.5.1 Discrete Uniform Distribution

The random experiment consisting of rolling one die can be modeled with a discrete uniform distribution where the random variable is the number of dots facing up. There are n values possible between integers a and b inclusive, and each value has a probability equal to $1/n$.

**Parameters:**

$$a, \quad b, \quad n = b - a + 1$$

**Support:**

Every integer between a and b, including a and b.

**PMF:**

$$P(Y = y) = 1/n$$

**CDF:**

$$F_Y(k) = \frac{\lfloor k \rfloor - a + 1}{b - a + 1}$$

**Mean:**

$$\mu_Y = \frac{a + b}{2}$$

**Variance:**

$$V\{Y\} = \sigma_Y^2 = \frac{n^2 - 1}{12}$$

### 1.5.2 Continuous Uniform Distribution

This is the distribution where a continuous variable has equal probability of taking values in any segment of equal width between a and b. Its pdf was given in Equation (1.8).

**Parameters:**

$a, \quad b$

**Support:**

Every real number between a and b, including a and b.

**PDF:**

$$f_Y(y) = 1/(b - a) \quad \forall \quad a \le y \le b, \quad 0 \quad \text{elsewhere}$$

**CDF:**

$$F_Y(y) = \begin{cases} 0 & \text{if } y \ < a \\ \frac{y-a}{b-a} & \text{if } \ a \leq y \leq b \\ 1 & \text{if } \ y > b \end{cases}$$

**Mean:**

$$\mu_Y = \frac{a+b}{2}$$

**Variance:**

$$V\{Y\} = \sigma_Y^2 = \frac{(b-a)^2}{12}$$

## 1.5.3 Binomial Distribution

This is the distribution of a random variable obtained by summing n independent Bernoulli random variables. The binomial random variable is the number of “successes” in a set of n independent trials with constant probability of success p. The folloiwng random variables are usually modeled with a Binomial distribution:

- Number of heads in n coin tosses.
- Number of seeds that germinate out of a set of 10.
- Number of bird nests found in a forest plot that actually has 5 nests.
- Number of salmon that survive and return to a hatchery stream out of 1000 young released.
- Number of apples with worms out of 20 selected randomly from a batch of 10000.

**Parameters:**

$$n = \text{no. of trials} \qquad p = \text{probability of success}$$

**Support:**

Natural numbers between 0 and n (it’s a discrete distribution).

**PMF:**

$$f_Y(r) = P(Y=r) = C_r^n \, p^r \, q^{n-r} \quad \forall \quad 0 \le r \le n, \quad 0 \quad \text{elsewhere}$$

**CDF:**

$$F_Y(k) = P(Y \le k) = \sum_{i=0}^{\lfloor k \rfloor} C_i^n \, p^i \, q^{n-i}$$

**Mean:**

$$\mu_Y = n \, p$$

**Variance:**

$$V\{Y\} = \sigma_Y^2 = n \, p \, (1-p)$$

In the binomial calculations, the combinations are combinations of positions where “success” happens. We use combinations because having “success” in positions 1 and 5 for example, is exactly the same as having success in position 5 and 1. Just note that the order of successes and failures matters but the order of the positions of the successes and failures does not! Coins tosses HHTT are not the same as HTHT. However, having H in positions 1 and 2 is the same as having H in positions 2 and 1.

There are four basic functions in R to do calculations reletd to the Binomial distribution:

- `dbinom(x, size, prob)` returns the value of the PMF.
- `pbinom(q, size, prob, lower.tail = TRUE)` returns the CDF left tail.
- `qbinom(p, size, prob, lower.tail = TRUE)` returns the quantile.

- `rbinom(n, size, prob)` returns random values with a Binomial distribution.

Analogous functions are available for several distributions in R. The functions are illustrated in the R chunk below.

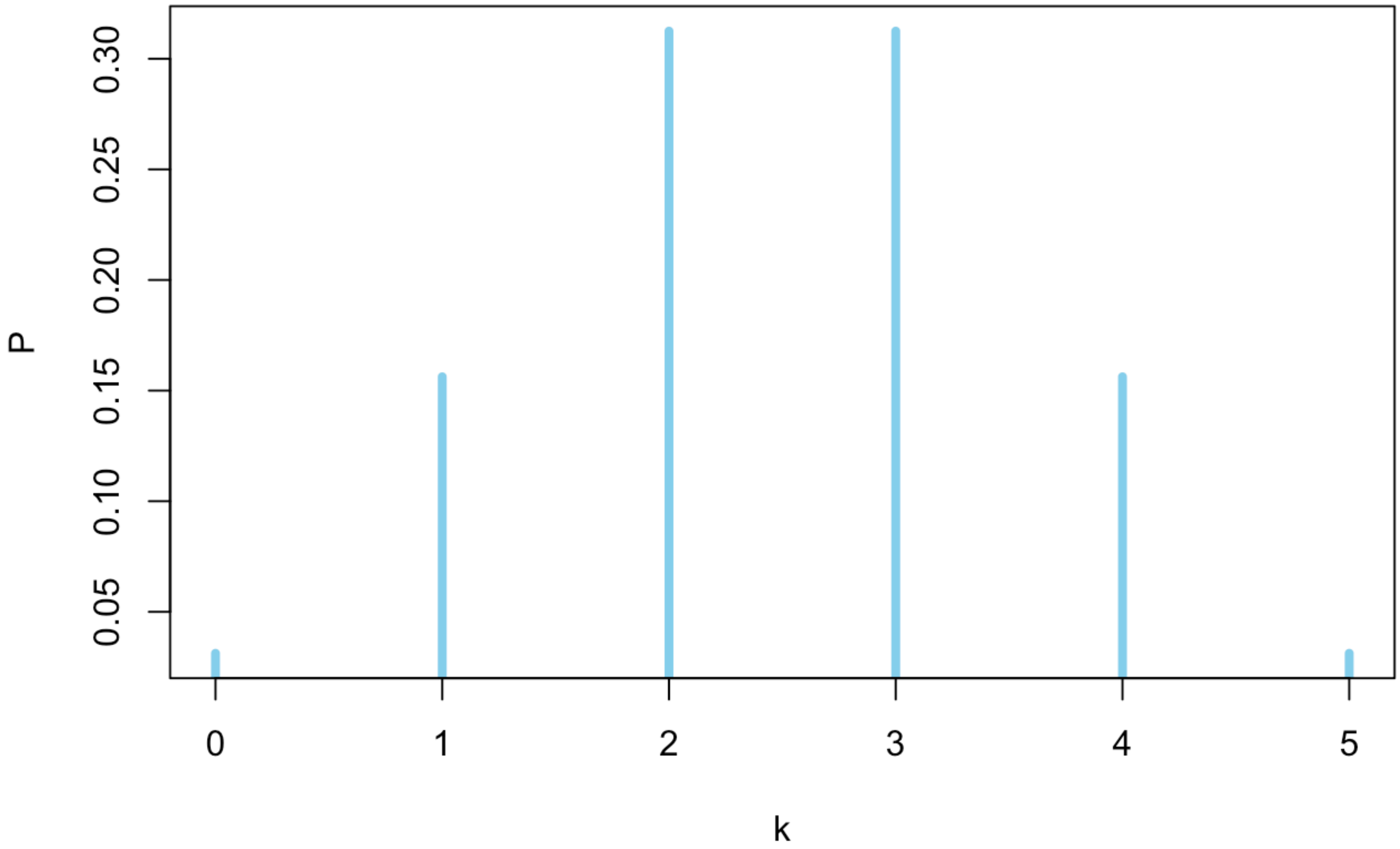
```r
# Probability of 3 heads in 5 coin tosses
dbinom(x = 3, size = 5, prob = 0.5)
```

```
## [1] 0.3125
```

```r
# Probability of k heads in 5 tosses
(df <- data.frame(k = 0:5, P = dbinom(x = 0:5, size = 5, prob = 0.5)))
```

```
##   k       P
## 1 0 0.03125
## 2 1 0.15625
## 3 2 0.31250
## 4 3 0.31250
## 5 4 0.15625
## 6 5 0.03125
```

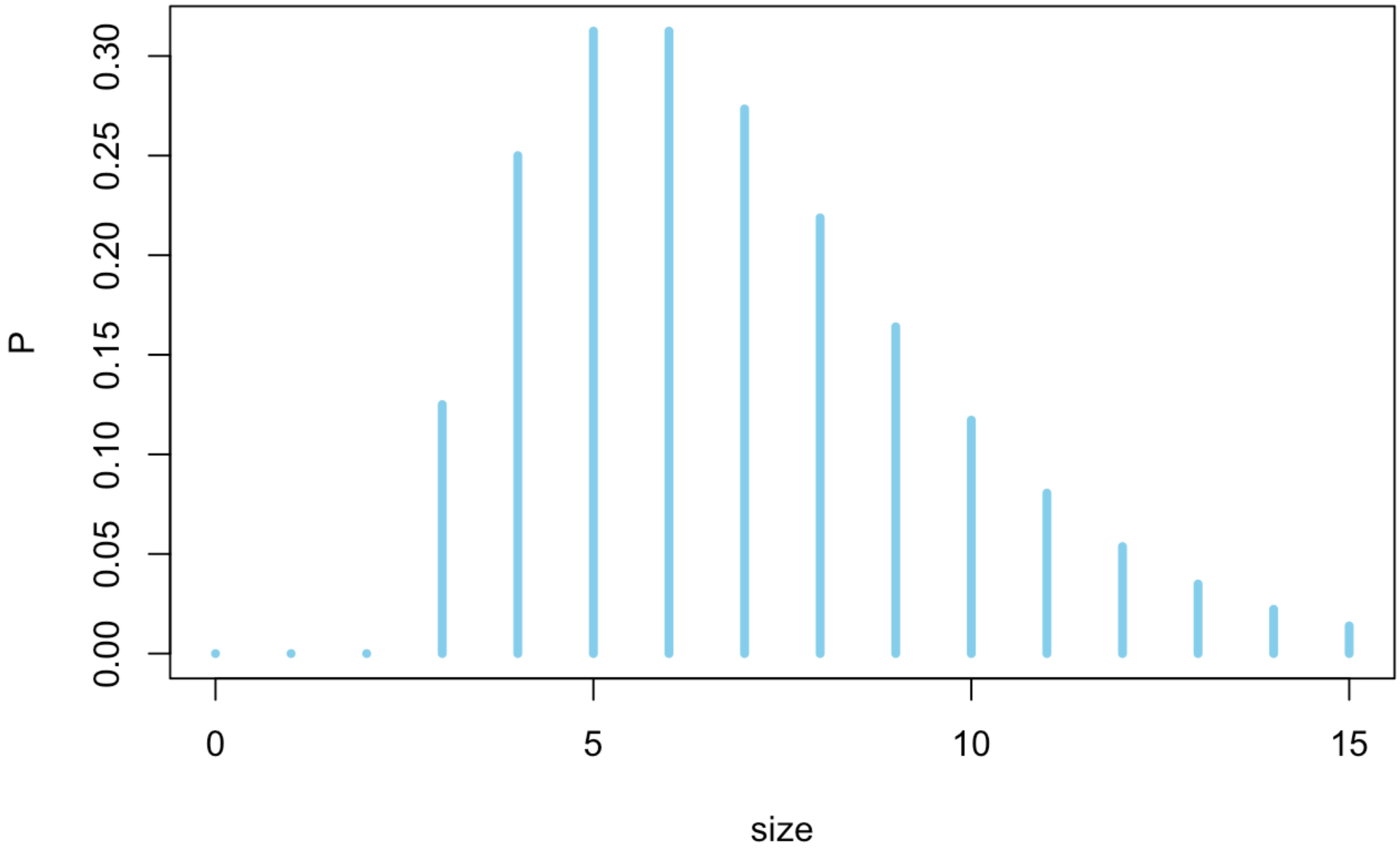
```r
plot(df, type = "h", col = "skyblue", lwd = 4)
```



```
# Probability of 3 heads as a function of number of tosses
(df <- data.frame(size = 0:15, P = dbinom(x = 3, size = 0:15, prob = 0.5)))
```

```
##    size             P
## 1     0 0.00000000000
## 2     1 0.00000000000
## 3     2 0.00000000000
## 4     3 0.12500000000
## 5     4 0.25000000000
## 6     5 0.31250000000
## 7     6 0.31250000000
## 8     7 0.27343750000
## 9     8 0.21875000000
## 10    9 0.16406250000
## 11   10 0.11718750000
## 12   11 0.08056640625
## 13   12 0.05371093750
## 14   13 0.03491210937
## 15   14 0.02221679688
## 16   15 0.01388549805
```

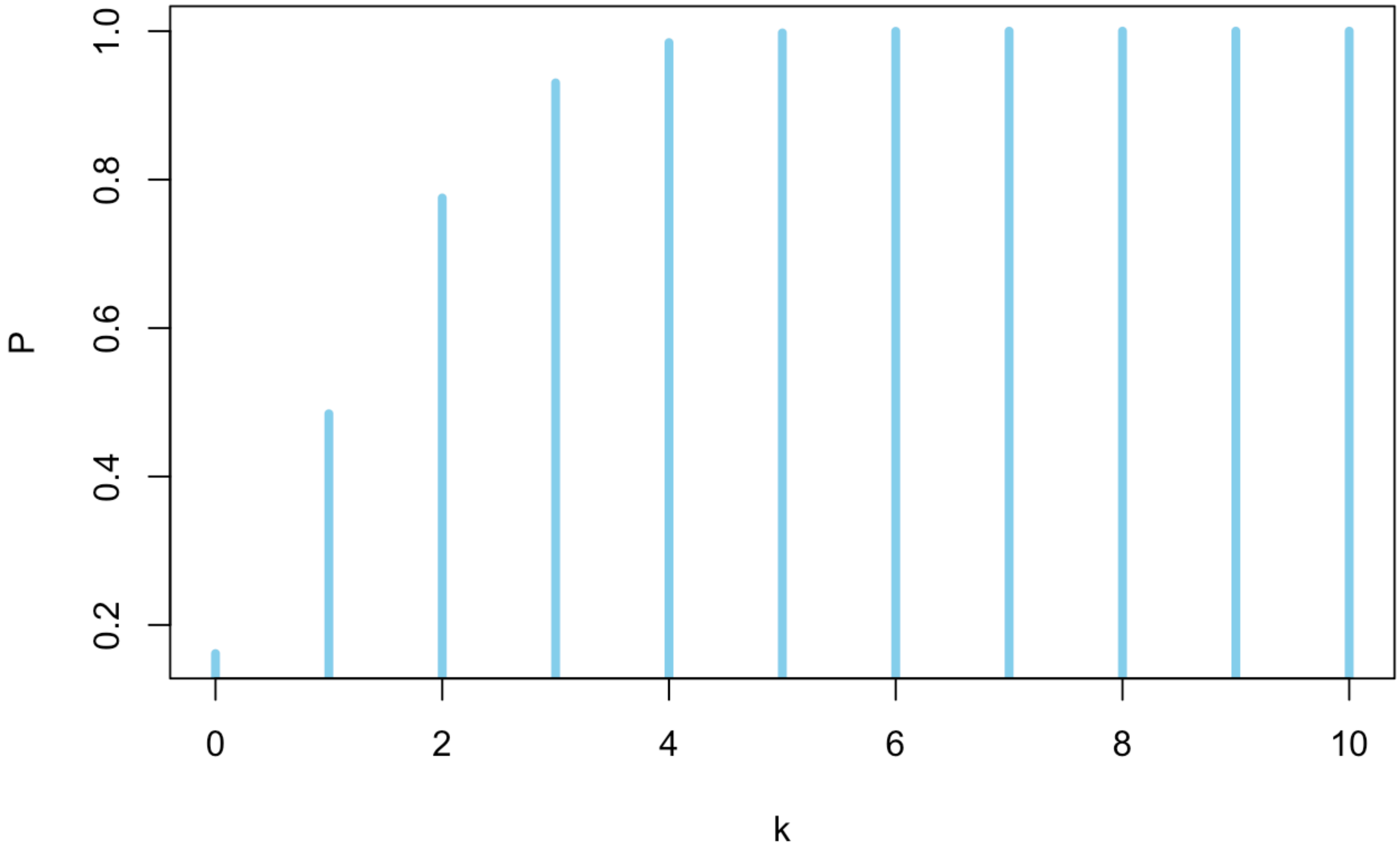
```
plot(df, type = "h", col = "skyblue", lwd = 4)
```



```r
# Cumulative distribution function for number of 5's in 10 die rolls
(df <- data.frame(k = 0:10, P = pbinom(q = 0:10, size = 10, prob = 1/6)))
```

```
##     k            P
## 1   0 0.1615055829
## 2   1 0.4845167487
## 3   2 0.7752267979
## 4   3 0.9302721574
## 5   4 0.9845380333
## 6   5 0.9975618435
## 7   6 0.9997324785
## 8   7 0.9999805511
## 9   8 0.9999991566
## 10  9 0.9999999835
## 11 10 1.0000000000
```

```
plot(df, type = "h", col = "skyblue", lwd = 4)
```



```r
# P(> 1) infected animals in a sample of 20 from a population where p = 0.02
# One way; note that q is included in the lower tail
pbinom(q = 2, size = 20, prob = 0.02, lower.tail = FALSE)
```

```
## [1] 0.007068693404
```

```r
# or another
1 - dbinom(x = 0, size = 20, prob = 0.02) -
    dbinom(x = 1, size = 20, prob = 0.02) -
    dbinom(x = 2, size = 20, prob = 0.02)
```

```
## [1] 0.007068693404
```

### 1.5.4 Poisson Distribution

The Poisson distribution is typically used to describe the number of events or occurrences that take place over periods of time or segments of space when the probability of event is constant over space and or time, and events are independent. The probability that an event happens in a specific segment of space or time does not depend on whether there are events on other segments. The following are examples of random variables that can be modeled with a Poisson distribution:

- Number of cars that pass by a corner every 10 minutes between 8 and 9 am every weekday.
- Number of birds observed in a forest plot during 1 hour.
- Number of plants of a given species of weed that are within each square meter in an alfalfa field.
- Number of water pumps that break down in California every minute.
- Number of times a machine is expected to fail in 5000 hours of operation.
- Number of weak points per 100 ft length of rope.

The parameter $\lambda$, which is the mean and the variance of the Poisson, is the mean number of events per unit space or time.

**Parameters:**

$$\lambda = \mu = \sigma^2$$

**Support:**

Natural numbers between 0 and $\infty$ (it's a discrete distribution).

**PMF:**

$$f_Y(y) = P(Y = y) = \frac{e^{-\mu} \, \mu^y}{y!} \quad \forall \quad 0 \leq y, \quad 0 \quad \text{elsewhere}$$

**CDF:**

$$F_Y(k) = P(Y \leq k) = e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$$

**Mean:**

$$\mu_Y = \lambda$$

**Variance:**

$$V\{Y\} = \sigma_Y^2 = \lambda$$

For the practical application of the Poisson distribution, if $\lambda$ is the mean rate of events per unit time or space, then the mean of the distribution is $\lambda \cdot s$, where $s$ is the size (length, duration, volume, etc.) of the interval considered. Suppose that the mean number of plants per unit area is 6 $m^{-2}$, what is the probability of observing more than 3 plants in a quadrat that is 0.5 m on each side?

The area of the quadrat is 0.25 $m^{-2}$, thus, the expected number of plants per quadrat is 6 $m^{-2}$ $\times$ 0.25 $m^2/quadrat$ = 1.5 $quadrat^{-1}$ or 1.5 plants per quadrat.

$$P(Y > 3) = 1 - P(Y \leq 2) = 1 - P(Y = 0) - P(Y = 1) - P(Y = 2)$$

$$= 1 - \frac{e^{-1.5}\,1.5^0}{0!} - \frac{e^{-1.5}\,1.5^1}{1!} - \frac{e^{-1.5}\,1.5^2}{2!}$$

$$= 1 - 0.2231 - 0.3347 - 0.251 = 0.1912$$

Obviously, the probability that there are more than 3 plants in a 1-$m^2$ quadrat is going to be larger than in a smaller quadrat. Both distributions are plotted in Figure 1.5

```
plot(0:20 + 0.05, dpois(x = 0:20, lambda = 1.5), type = "h", lwd = 4, col = "o
points(0:20 - 0.05, dpois(x = 0:20, lambda = 6), type = "h", lwd = 4, col = "b
```

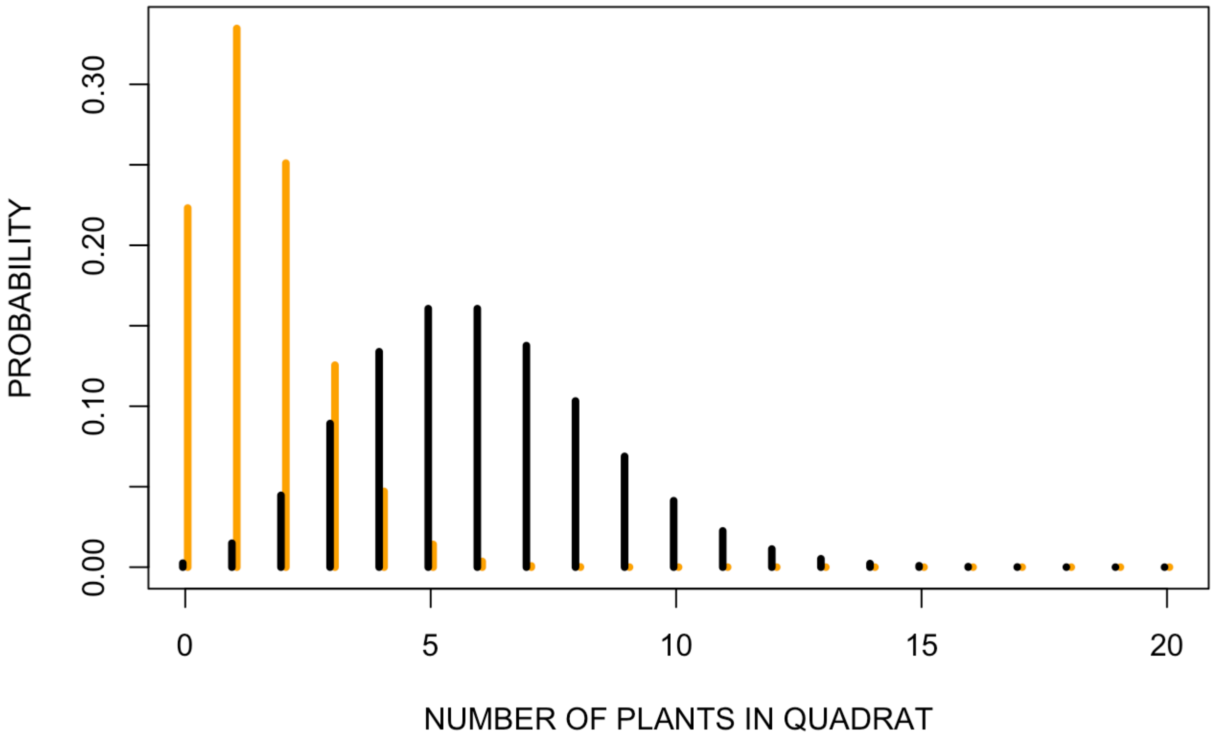


Figure 1.5: Poisson distributions for number of plants per quadrat in a population where the average number of plants per square meter is 6. Both distributions have probabilities that are positive for any positive integer, the distributions are truncated at values of Y for which the probability is too small to be seen in the plot. Abscissa values have been slightly offset so both distributions can be seen without overlap.

Note that when asked for the probability of values greater than something in a Poisson, you have to use the probability of the complement, otherwise you would have to sum an infinite number of terms.

There are four basic functions in R to do calculations reletd to the Poisson distribution:

- `dpois(x, lambda)` returns the value of the PMF.
- `ppois(q, lambda, lower.tail = TRUE)` returns the CDF left tail.
- `qpois(p, lambda, lower.tail = TRUE)` returns the quantile.

- `rpois(n, lambda)` returns random values with a Poisson distribution.

The functions are illustrated in the R chunk below. The number of crossovers that happen in a region of a chromosome during meiosis can be modeled with a Poisson distribution. Suppose that we consider a region where the mean crossover rate is 0.8, and that each meiosis represents an independent event.

```r
# Probability of 0 crossovers
dpois(x = 0, lambda = 0.8)
```

```
## [1] 0.4493289641
```

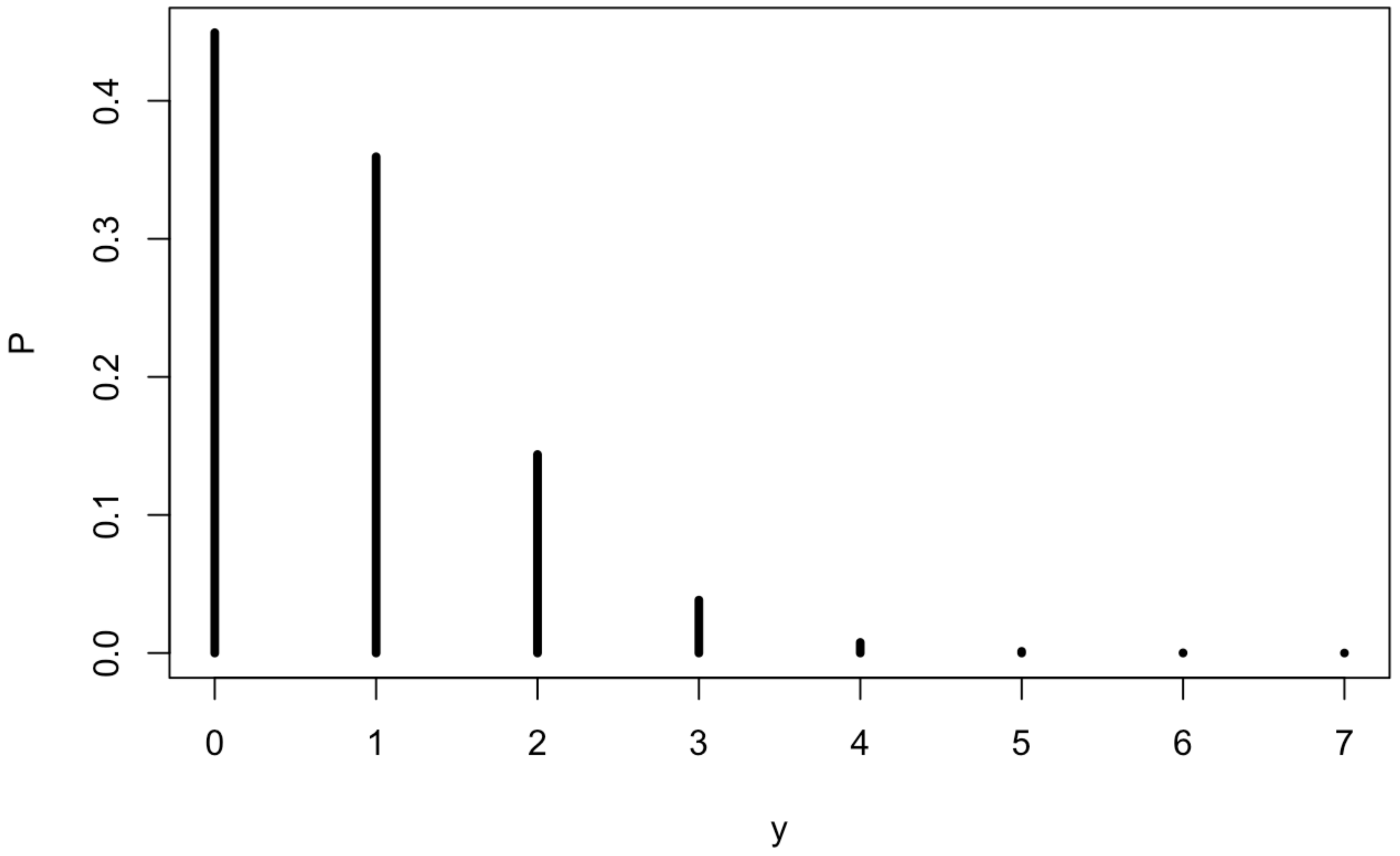
```r
# Probability of 2 crossovers
dpois(x = 2, lambda = 0.8)
```

```
## [1] 0.1437852685
```

```r
# Probability of y crossovers
(df <- data.frame(y = 0:7, P = dpois(x = 0:7, lambda = 0.8)))
```

```
##   y                P
## 1 0 0.44932896411722
## 2 1 0.35946317129378
## 3 2 0.14378526851751
## 4 3 0.03834273827134
## 5 4 0.00766854765427
## 6 5 0.00122696762468
## 7 6 0.00016359568329
## 8 7 0.00001869664952
```

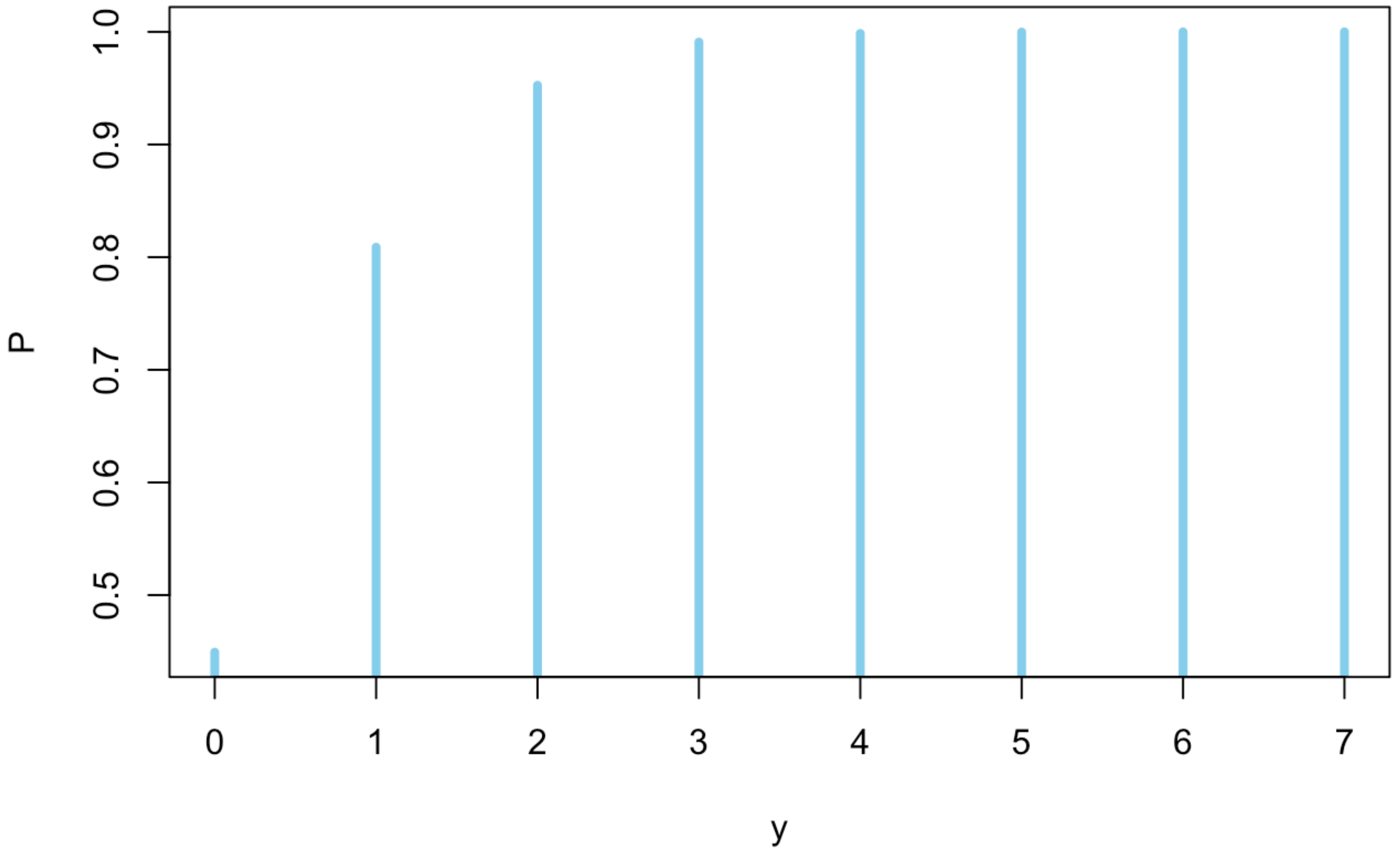
```r
plot(df, type = "h", lwd = 4)
```



```r
# Cumulative distribution function for number of crossovers
(df <- data.frame(y = 0:7, P = ppois(q = 0:7, lambda = 0.8)))
```

```
##   y            P
## 1 0 0.4493289641
## 2 1 0.8087921354
## 3 2 0.9525774039
## 4 3 0.9909201422
## 5 4 0.9985886899
## 6 5 0.9998156575
## 7 6 0.9999792532
## 8 7 0.9999979498
```

```
plot(df, type = "h", col = "skyblue", lwd = 4)
```



```r
# Expected proportion of gametes with 3 or more crossovers
# One way; note that q is included in the lower tail
ppois(q = 3, lambda = 0.8, lower.tail = FALSE)
```

```
## [1] 0.0090798578
```

```r
# or another
1 - dpois(x = 0, lambda = 0.8) -
    dpois(x = 1, lambda = 0.8) -
    dpois(x = 2, lambda = 0.8) -
    dpois(x = 3, lambda = 0.8)
```

```
## [1] 0.0090798578
```

### 1.5.5 Normal Distribution

The Normal or Gaussian distribution is most important in statistics and science, because it is an accurate model for a multitude of variables measured. As sample size increases, many estimated parameters tend to have a normal distribution, regardless of the original distribution from where the sample is obtained. The normal distribution describes difussion of gas particles in a volume, noise in electromagnetic signal, values in chaotic systems, and all sorts of variables measured in the empirical sciences.

For any given variance and mean, the Normal distribution is the continous distribution with maximum entropy. This means that when one knows the mean and variance of a continuous distribution but nothing else about it, assuming that it has a Normal distribution is the choice that imposes or adds the least prior information.

**Parameters:**

$$\mu \text{ and } \sigma^2$$

**Support:**

Real numbers between $-\infty$ and $\infty$ (it’s a continuous distribution).

**PMF:**

$$f_Y(y) = P(Y = y) = \frac{1}{2\,\pi\,\sigma^2}\; e^{-\frac{(y-\mu)^2}{2\,\sigma^2}} =$$

$$= \frac{1}{2\,\pi\,\sigma^2}\; exp\left[-0.5\left(\frac{y-\mu}{\sigma}\right)^2\right]$$

Note that the probability density depends on the square of the deviation from the mean. As a result, the Normal distribution is symmetric about the mean, the height of the curve at $y = \mu + a$ is equal to the height at $y = \mu - a$: $f_Y(\mu + a) = f_Y(\mu - a)$

**CDF:**

$$\Phi(y) = \int_{-\infty}^{y} \frac{1}{2\,\pi\,\sigma^2}\; exp\left[-0.5\left(\frac{y-\mu}{\sigma}\right)^2\right]\; dy$$

The CDF of the Normal distribution is a well defined function but it cannot be written down with other elementary functions. We use the Greek letter $\Phi$ to refer to it.

**Mean:**

$$\mu_Y$$

**Variance:**

$$V\{Y\} = \sigma_Y^2$$

Although the Normal pdf has positive values for all real numbers, we frequently use it as a approximation for random variables that are strictly positive such as mass, length, and other biophysical quantities that cannot be negative. For the approximation to be a good one, the mean has to be at least 4 times the standard deviation. This way, less than 1/30,000 of the distribution is expected to be below zero, making errors of little practical consequence. In cases when the mean is closer to zero than 3-4 standard deviations, we should consider other distributions that have the positive reak numbers as support, including 0. The log-normal and members of the Gamma family of distributions would be good choices.

### Calculations with the Normal distribution

Several of the calculations we present here for the Normal distribution can be performed with any distribution, particularly with symmetric continuous distributions. Most of the time we are interested in areas under the Normal pdf $f(y)$ between values of the random variable, which is the same as differences in the height of $\Phi(y)$ at the values of Y.

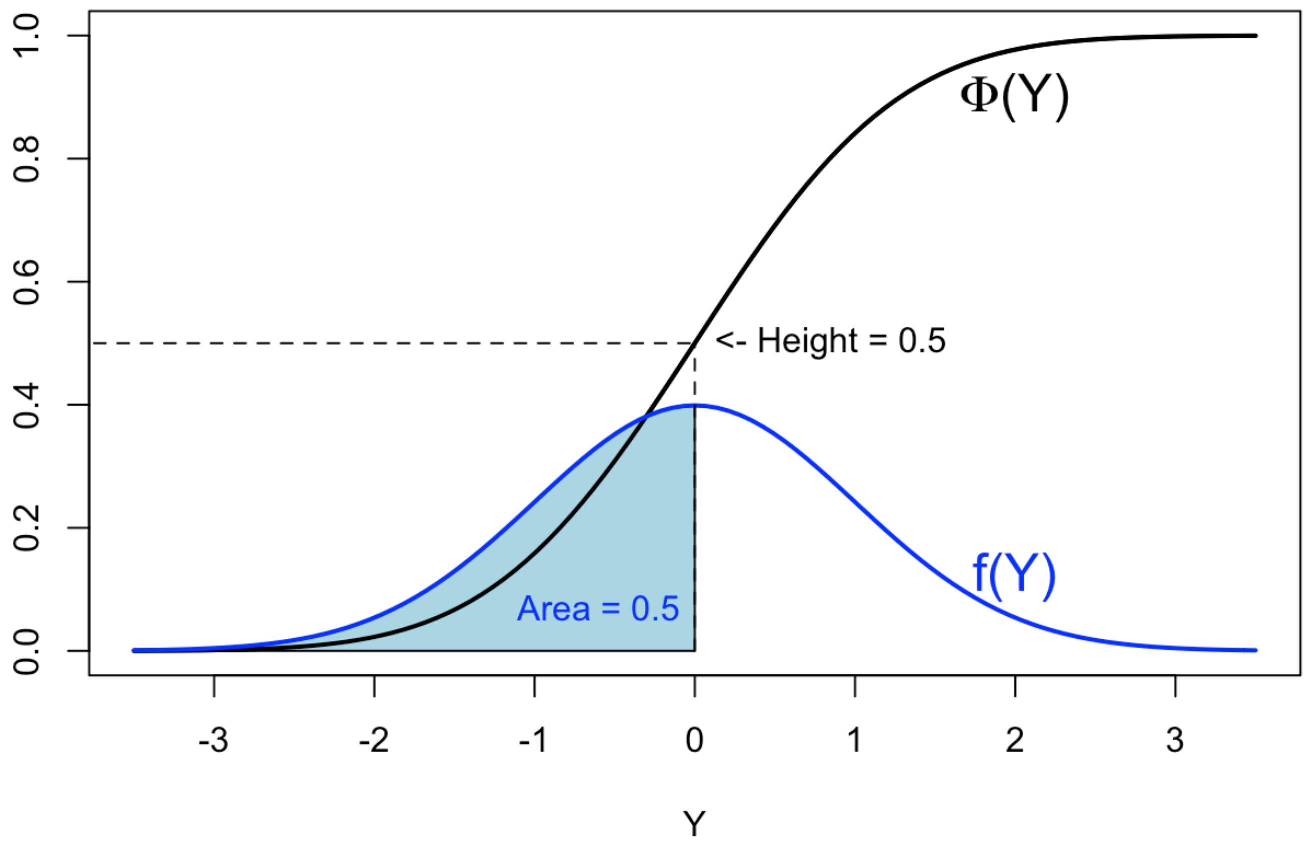


Figure 1.6: Probability density function (blue) and cumulative distribution function (black) for the standard Normal distribution. The shaded area under the pdf is te probability that the random variable Y takes values less than 0, which is the height of the cumulative distribution function at Y = 0.

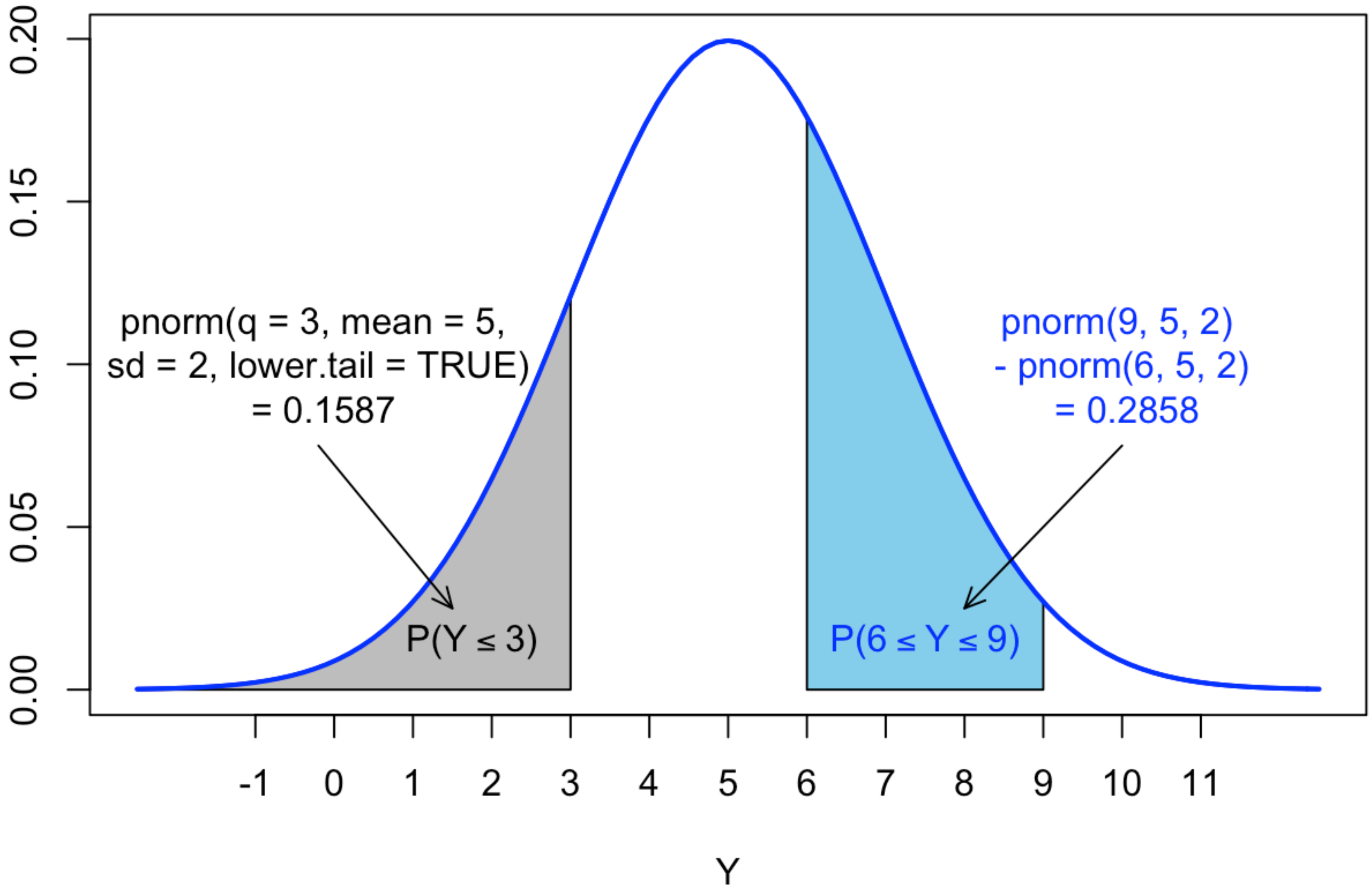
Any Normal RV can be mapped to the Standard Normal Distribution (SND) by the linear transformation consisting in subtracting the mean and dividing by the standard deviation.

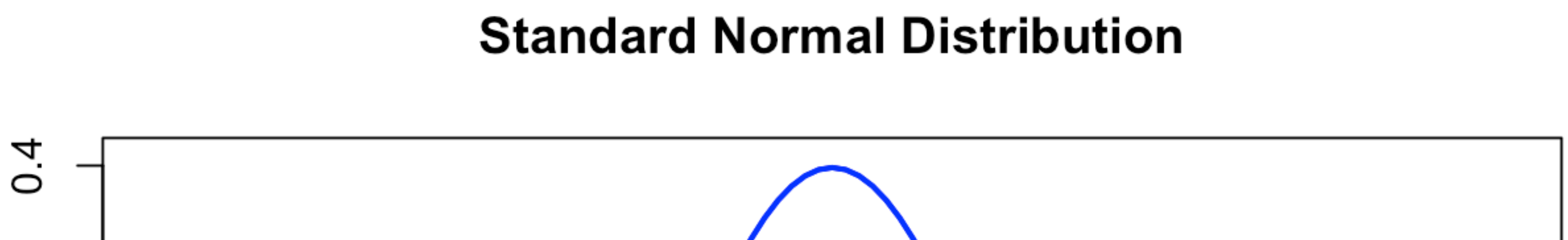
$$P(Y \leq y) = P\left(Z = \frac{Y-\mu}{\sigma} \leq \frac{y-\mu}{\sigma} = z\right)$$

If Y has a Normal distribution with mean $\mu$ and variance $\sigma^2$ and Z has a SND, which we write as $Z \sim N(0,1)$, then

$$P(y_1 \leq Y \leq y_2) = P\left(\frac{y_1-\mu}{\sigma} \leq Z \leq \frac{y_2-\mu}{\sigma}\right)$$

What is the probability that $Y \sim N(5,4)$ takes values between 6 and 9, or that it takes values lower than3? Those probabilities are represented as areas under the curve of $Y \sim N(5,4)$ and under the SND $Z \sim N(0,1)$ in Figure 1.7.





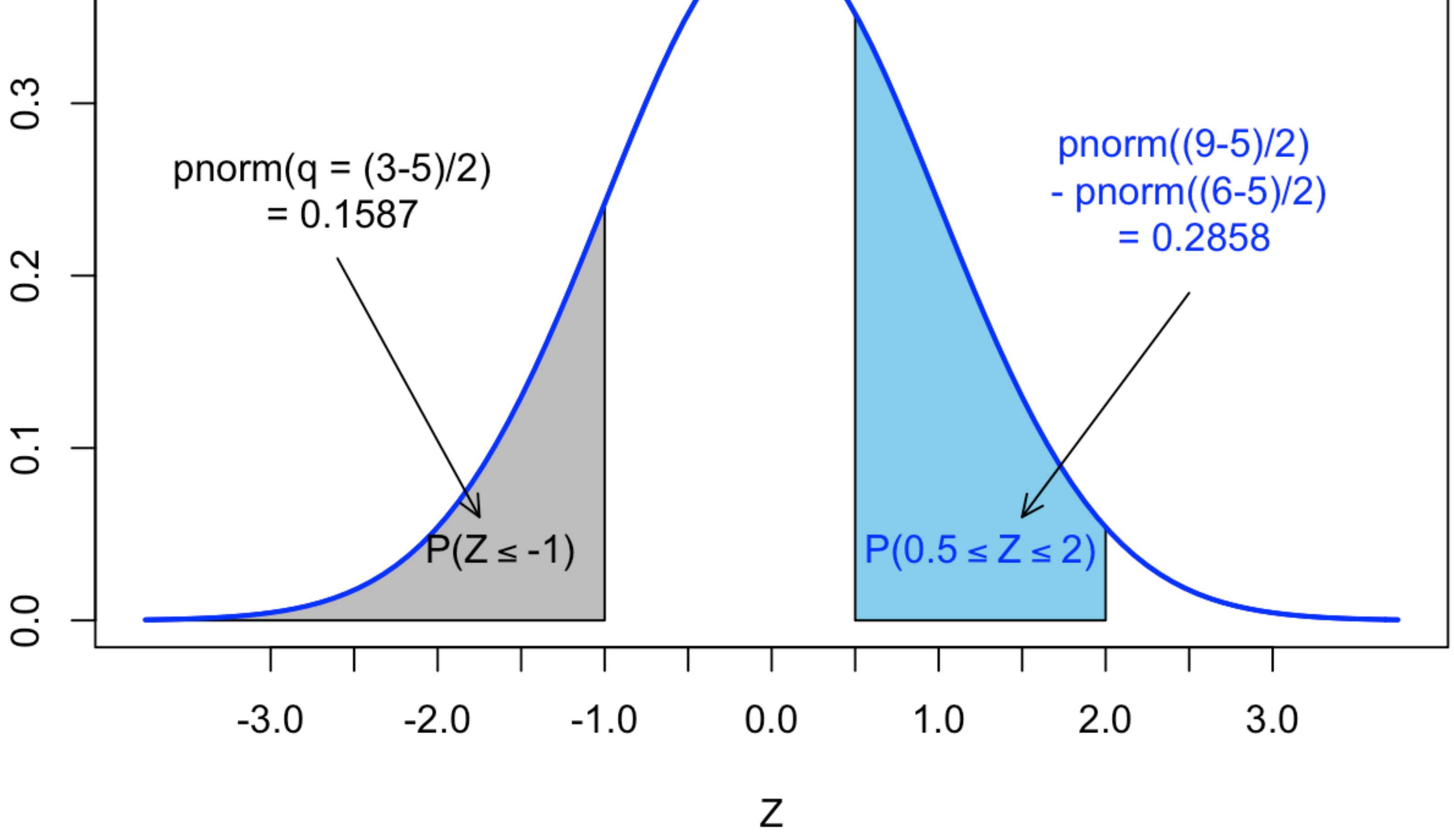


Figure 1.7: Equivalence of areas under curve in a normal distribution with mean 5 and standard deviation 4, and the standard Normal Distribution.

The areas under the curve are used to represent proportions of the population with values between the specified extremes. Imagine that the length of Hass avocados received for sorting and packing at a plant has has a normal distribution with mean 9 cm and variance 1.21.cm$^2$. The following are questions that we can answer using the normal distribution (data and situations described are fictitious).

1. A buyer wants avocados that are longer than 8 cm. What proportion of the avocados received could go to that buyer?

We are given a set value of the random variable Y and are asked to find a proportion, which we interpret as a probability. The set value $y = 8$ is called a **quantile**. We look up the value of the quantile in the graph and then calculate the probability. Because the total ara under the curve is 1.0, the answer can be calculated in two different ways. The different ways were more relevant when we had to use printed tables to find quantiles and probabilities, because tables usually included only one half of the distribution. The other half is the mirror image and can be derived as such.

$$Y \sim N(9, 1.21) \implies P(Y \leq 8) = 1 - P(Y \geq 8)$$

The R function to go from normal quantiles to probabilities is `pnorm` and it corresponds to the cumulative function $\Phi$.

```
# probability that Y is less than 8

1 - pnorm(q = 8, mean = 9, sd = sqrt(1.21), lower.tail = TRUE)
```

```
## [1] 0.8183489296
```

```
pnorm(q = 8, mean = 9, sd = sqrt(1.21), lower.tail = FALSE)
```

```
## [1] 0.8183489296
```

```
# The lower tail and standard normal distribution are the defaults.

1- pnorm(q = (8-9)/sqrt(1.21))
```

```
## [1] 0.8183489296
```

About 82% of the avocados can be packed for the buyer.

2. The plant wants to pack boxes that sell for a low price that requires that no more than 12.5% of the avocados be rejected. What is the shortest avocado that will be packed?

This is the reverse of the previous question; we are given a percentage or proportion and asked for a value of Y or quantile. What is the quantile of $Y \sim N(9, 1.21)$ that corresponds to a probability of 0.125? The R function to go from probability to quantile is ‘qnorm’ and it corresponds to the inverse of $\Phi$.

```r
# 0.125 quantile or value of Y below which is 0.125 of the area under curve

qnorm(p = 0.125, mean = 9, sd = sqrt(1.21), lower.tail = TRUE)
```

```
## [1] 7.734615682
```

```r
# Is the same as the value that leaves 1 - 0.125 probability above it.

qnorm(p = 1 - 0.125, mean = 9, sd = sqrt(1.21), lower.tail = FALSE)
```

```
## [1] 7.734615682
```

```r
# Standard normal distribution is the default.

qnorm(p = 0.125) * sqrt(1.21) + 9
```

```
## [1] 7.734615682
```

In the last way to calculate the quantile sought we used the reverse of the definition of a standardized variable to get the quantile in the units of avocado length:

$$\Phi^{-1}(p = 0.125) = Z_{0.125} = \frac{Y_{0.125} - \mu_Y}{\sigma_Y}$$

$$\implies \; Y_{0.125} = Z_{0.125} \times \sigma_Y \; + \; \mu_Y$$

3. Avocados prices are $1.00 for those shorter than 8 cm, $1.50 between 8 and 9.5 cm, $2.00 between 9.5 and 11 cm and $1.25 for those longer ones. What is the expected price of a randomly selected avocado? What is the variance of the price per avocado?

The price of an avocado is a function of its length, and the function can be simply represented with a table, as there is no “formula” given. Therefore, the normal distribution is used only in the first step, to calculate the probability mass function of the resulting discrete distribution.

```r
# proportion small < 8 cm
pS <- pnorm(8, mean = 9, sd = sqrt(1.21))

# proportion small < 8 cm
pM <- pnorm(9.5, mean = 9, sd = sqrt(1.21)) - pS

# proportion small < 8 cm
pL <- pnorm(11, mean = 9, sd = sqrt(1.21)) - pS - pM

# proportion small < 8 cm
pXL <-1 - pS - pM - pL

(PMF <- data.frame(
   price = c(1, 1.5, 2, 1.25),
   Prob = c(pS, pM, pL, pXL)
   )
)
```

```
##   price         Prob
## 1  1.00 0.1816510704
## 2  1.50 0.4936307877
## 3  2.00 0.2901999679
## 4  1.25 0.0345181740
```

```r
# Check calculation of probabilities

sum(PMF$Prob) # should be exactly 1
```

```
## [1] 1
```

```
# Apply definition of expectation of discrete distribution

(meanPrice = sum(PMF$price * PMF$Prob))
```

```
## [1] 1.545644905
```

- When you are asked for a probability, for continous RV’s use the R function starting with “p”, as in `pnorm`.
- When you are asked for a critical value or a value of the random variable, use the R function that starts with “q” as in `qt`.

### 1.5.6 $\chi^2$ Distribution

This distribution is the distribution of a random variable resulting from summing the squares of multiple independent SND’s. The $\chi^2$ distribution has one parameter, the number of independent SND squared and summed, which is called the *degrees of freedom* of the distribution.

$$Y = \sum_{i=1}^{n} Z_i^2$$

$$\text{where} \quad Z_i \sim N(0,1) \quad \forall\, i$$

$$\implies \quad Y \sim \chi_n^2 = \chi^2(n)$$

This distribution is used to do hypothesis testing and confidence intervals for estimated variances, as well as for assessing independence in table of events and comparing observed to expected frequencies.

**Parameters:**

$$\nu \quad \text{degrees of freedom}$$

**Support:**

Real numbers between $0$ and $\infty$ (it's a continuous distribution).

**PMF:**

Both the PMF and CDF involve gamma functions, which are not in the typical toolbox. These are extremely useful functions, but they are used in the background. They are mention here just as a general math culture point, but we will not use or study them further.

$$f_Y(y) = P(Y = y) = \frac{y^{\nu/2-1}\, e^{-y/2}}{2^{\nu/2}\, \Gamma(\nu/2)} \qquad y > 0$$

$\Gamma()$ is the "Gamma" function, which is an extension of the factorial to real and complex numbers. When $\nu$ is a strictly positive integer, $\Gamma(\nu) = (\nu - 1)!$

**CDF:**

$$F_Y(y) = P(Y \leq y) = \frac{\gamma(\nu/2,\, y/2)}{\Gamma(\nu/2)} \qquad y > 0$$

where $\gamma(\nu/2,\, y/2)$ is the *lower incomplete gamma function*.

**Mean:**

$$E\{Y\} = \mu_Y = \nu$$

**Variance:**

$$V\{Y\} = 2\,\nu$$

R has four basic functions dealing with $\chi^2$, just like for the other common distributions. R help [@R-base] states:

`dchisq(x, df, ncp = 0, log = FALSE)` `pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)` `qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)` `rchisq(n, df, ncp = 0)`

where the arguments are

- x, q: vector of quantiles.
- p: vector of probabilities.
- n: number of observations. If length(n) > 1, the length is taken to be the number required.
- df: degrees of freedom (non-negative, but can be non-integer).
- ncp: non-centrality parameter (non-negative).
- log.p: logical; if TRUE, probabilities p are given as log(p).
- lower.tail: logical; if TRUE , probabilities are P[X ≤ x], otherwise, P[X > x].

The non-centrality parameter extends the distribution to represent the sum of normal RV's whose means are not zero, but where all have variance 1. The ncp is the sum of the squared means.

As the number of degrees of freedom increase, $\chi^2$ tends to a normal distribution.

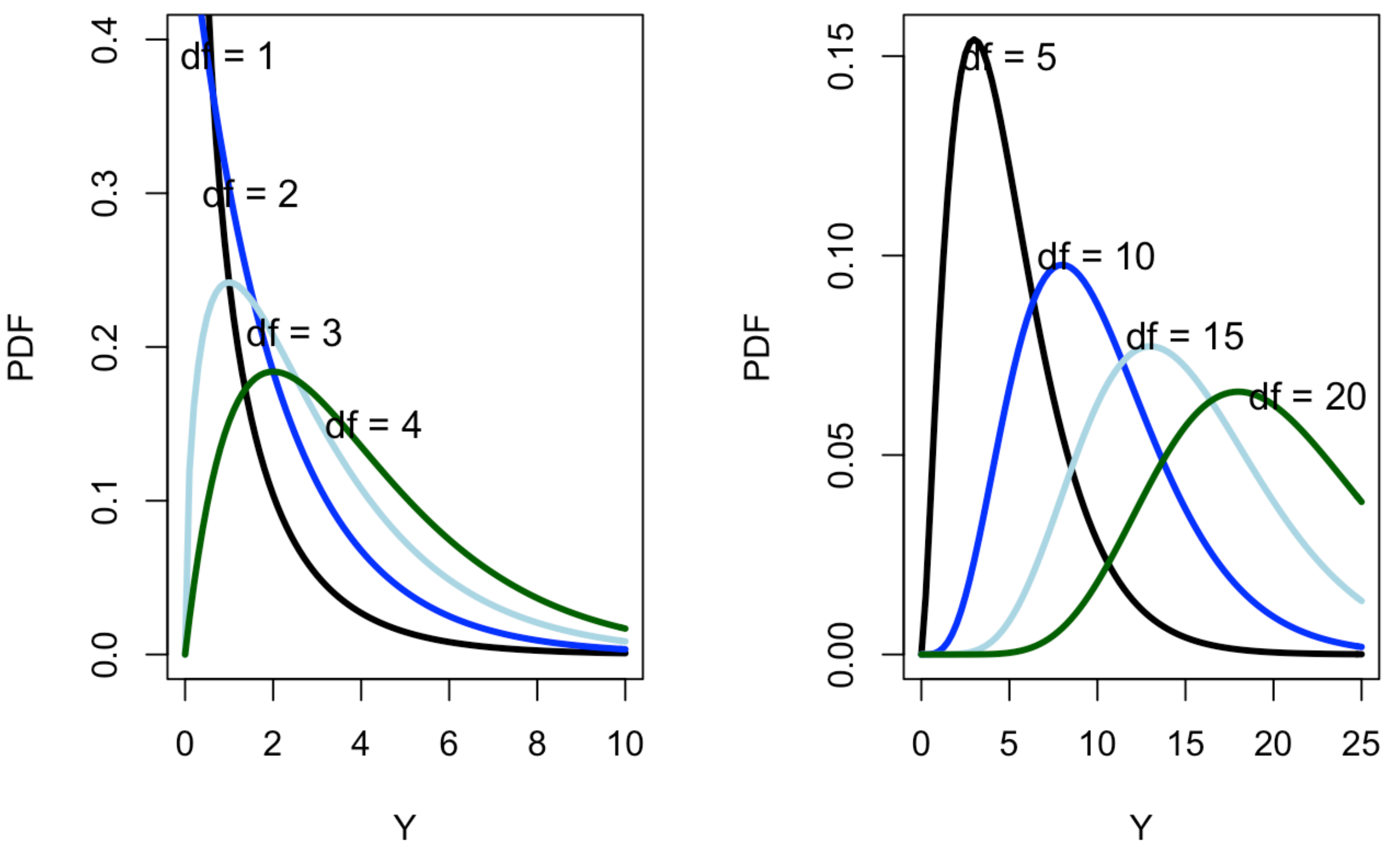


Figure 1.8: Probability density functions for $chi^2$ distributions with increasing degrees of freedom.

### 1.5.7 Student's t Distribution

This distribution is perhaps the second most used one after the SND. t-distributed random variables result from dividing a standard Normal by the square root of the ratio of an independent $\chi^2$ to its degrees of freedom.

$$\text{If} \quad Z \sim N(0,1) \quad \text{and} \quad V \sim \chi^2(\nu) \quad \text{are independent}$$

$$\implies T = \frac{Z}{\sqrt{V/\nu}} \sim t(\nu)$$

The t distribution is symmetric and approaches the SND as as degrees of freedom increase. For degrees of freedom less than 30, the PDF of the t distribution has more variance, or "heavier"tails than the normal. We can think of the additional variance comes from the uncertainty due to the independent variable V in the denominator of T.

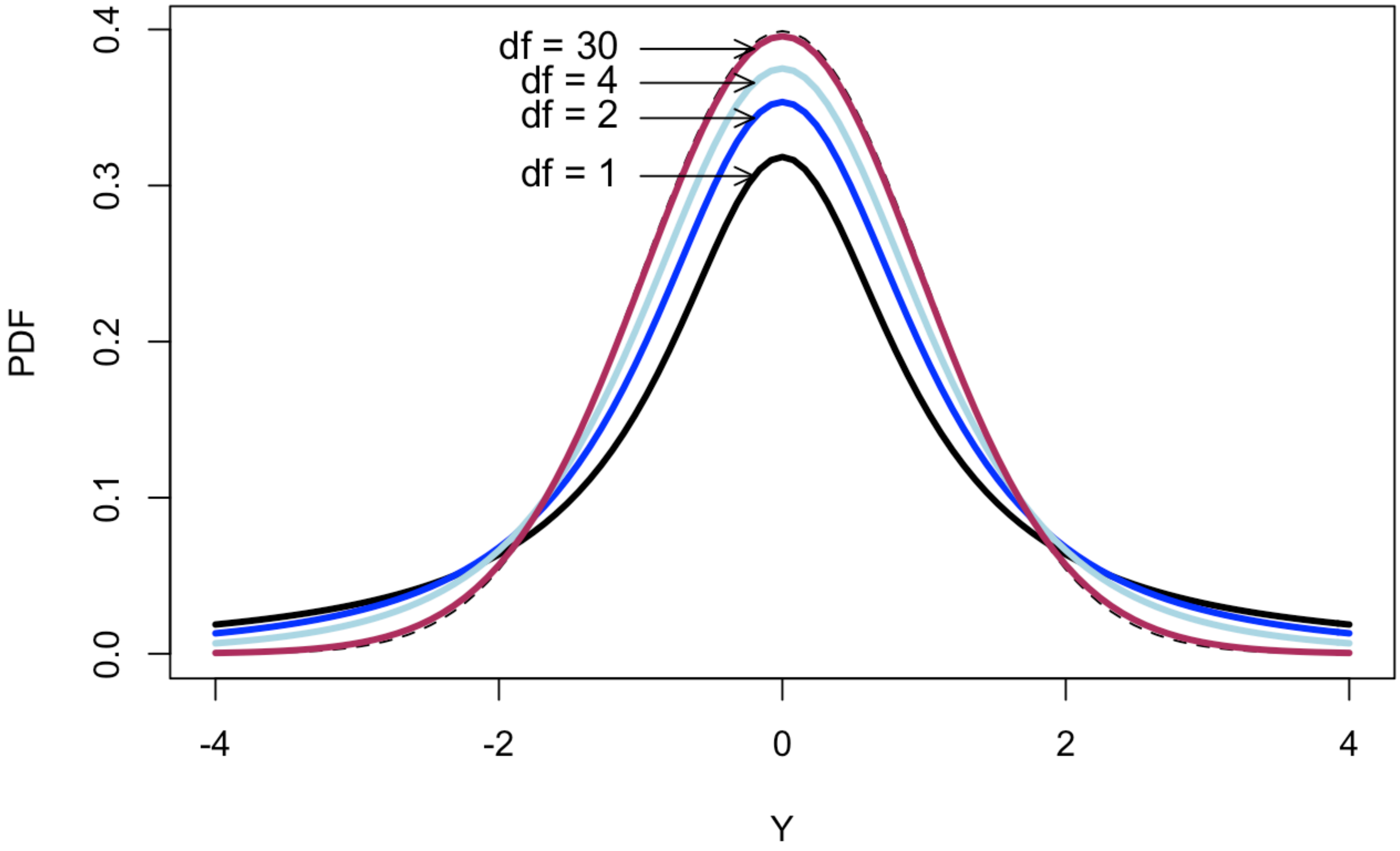


Figure 1.9: Comparison of t distributions with increasing degrees of freedom and the standard Normal distribution. The SND is drawn with the thin dashed line.

What are the “critical” t values that leave 5% of the population in two equal tails?

```
knitr::kable(data.frame(
   df = 1:30,
   t0.025 = qt(p = 0.025, df = c(1:30)),
   t0.975 = qt(p = 0.975, df = c(1:30))),
   caption = "Critical t values for 95% confidence with two tails.") %>%
   kable_styling(full_width = FALSE)
```

Table 1.4: Critical t values for 95% confidence with two tails.

| df | t0.025 | t0.975 |
|---|---|---|

|  |  |  |
|---:|---:|---:|
| 1 | -12.706204736 | 12.706204736 |
| 2 | -4.302652730 | 4.302652730 |
| 3 | -3.182446305 | 3.182446305 |
| 4 | -2.776445105 | 2.776445105 |
| 5 | -2.570581836 | 2.570581836 |
| 6 | -2.446911851 | 2.446911851 |
| 7 | -2.364624252 | 2.364624252 |
| 8 | -2.306004135 | 2.306004135 |
| 9 | -2.262157163 | 2.262157163 |
| 10 | -2.228138852 | 2.228138852 |
| 11 | -2.200985160 | 2.200985160 |
| 12 | -2.178812830 | 2.178812830 |
| 13 | -2.160368657 | 2.160368657 |
| 14 | -2.144786688 | 2.144786688 |
| 15 | -2.131449546 | 2.131449546 |
| 16 | -2.119905299 | 2.119905299 |
| 17 | -2.109815578 | 2.109815578 |
| 18 | -2.100922040 | 2.100922040 |
| 19 | -2.093024054 | 2.093024054 |
| 20 | -2.085963447 | 2.085963447 |
| 21 | -2.079613845 | 2.079613845 |
| 22 | -2.073873068 | 2.073873068 |
| 23 | -2.068657610 | 2.068657610 |

| | | |
|---|---|---|
| 24 | -2.063898562 | 2.063898562 |
| 25 | -2.059538553 | 2.059538553 |
| 26 | -2.055529439 | 2.055529439 |
| 27 | -2.051830516 | 2.051830516 |
| 28 | -2.048407142 | 2.048407142 |
| 29 | -2.045229642 | 2.045229642 |
| 30 | -2.042272456 | 2.042272456 |

### 1.5.8 F Distribution

This distribution is the workhorse of hypothesis testing when using analysis of variance. The distribution arises when one estimated variance is divide by an independent estimate of the variance. More specifically, if random variables $U_1$ and $U_2$ have $\chi^2$ distributions with degrees of freedom $\nu_1$ and $\nu_2$, then

$$Y = \frac{U_1/\nu_1}{U_2/\nu_2} \sim \mathcal{F}_{\nu_1,\nu_2}$$

As we will see in a later chapter, if treatments means are not different, the quotient of mean squares in analysis of variance is supposed to have an $\mathcal{F}$ distribuion.

The PDF and CDF for the $\mathcal{F}$ distribuion are complicated and not presented here.

**Parameters:**

$$\nu_1 \quad \text{degrees of freedom of the numerator}$$

$$\nu_2 \quad \text{degrees of freedom of the denominator}$$

**Support:**

Real numbers between $0$ and $\infty$ (it’s a continuous distribution).

**Mean:**

$$E\{Y\} = \frac{\nu_2}{\nu_2 - 2} \quad \text{defined only for } \nu_2 > 2$$

R has four basic functions dealing with $\mathcal{F}$, just like for the other common distributions.

`df(x, df1, df2, ncp, log = FALSE)` `pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)` `qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)` `rf(n, df1, df2, ncp)`

- To get information about all the distributions implemented in the `stast` package of R, run `help("Distributions")` or type "Distributions" in the search field of the Help tab.

## 1.6 Central Limit Theorem

## 1.7 Sampling and samples

Applied statistics for plant, animal and environmental sciences deals almost exclusively with making probabilistic statements about unknown and unknowlable population parameters and functions of those parameters. We make estimations and give objective measure of the certainty we have about the statements. Instead of simply calculating probabilities for events or values taken by realizations of random variables with known distributions, as emphasized in the chapter about probability and in the introduction of random variables, we take samples and derive data that are realizations of random variables of unknown distributions and guess the characteristics of the populations where the samples cames from.

### 1.7.1 Universe, population and sample

**Universe** Refers to the collection of objects or units in the population. For example, all people enrolled in PLS120 constitute a universe of interest to the instructor.

**Population** Is the set of all values of a specific characteristics of all units in the universe. For example, the number of electronic devices owned by each student, or the interpupillary distance of all students in PLS120. Other populations could be the number of days between ovulations in dairy cattle in the US, where “dairy cattle in the US” means either all dairy cattle individuals in the US at some point in time (it is chaging constantly) or all animals that came and will come from a theoretical population of “eternal dairy cattle.”

**Sample** A set of students from the PLS120 universe is a sample, and the values of their characteristics (number of devices, etc.) are samples of the corresponding populations.

Frequently, we use the terms population and universe as if they were the same, and this does not cause much confusion. But keep in mind that a unverse of objects can have as many populations as variables we can define in that universe. The distincion is relevant because we will have to use different distribution functions for the different populations. The number of electronic devices might be modeled with a Poisson distribution, but not with a normal distribution, because it is discrete and has a mean that is close to zero.

- A **sample** is a set of population units or values obtained from a statistical population by a well defined procedure.

The link between population characteristics such as its parameter values and the results of any calculations done with sample values is strictly dependent on the sampling procedure. There are many well defined sampling procedures that were designed for specific situations, including simple random sampling, cluster sampling, stratified sampling and double sampling. For a good introduction to sampling theory and methods, see [@Thompson1992]. In this course we will use almost exclusively random sampling with replacement, but some other methods are briefly introduced.

### 1.7.2 No Representative samples allowed

You may have noticed that in the exercises and examples about probability and distributions we always stated (or should have!) "… a randomly selected …" or something like that. We did no state things like "… a representative unit from …" or "… selected based on how much I liked them …" The random selection of the observed units is what allowed us to use the distribution being considered to make statements about the probabilities associated with different values of the random variable. In particular, we never stated that some of the possible values of the random variable would be excluded. For example, we never stated things like "We flip a coin 10 times, but if we get 10 heads we ignore it and try again." The experiments were random and no elements in the sample space were excluded from the possible results.

- The assignment of probabilities to events and random variables is strictly dependent on the details of the random experiment. We deal exclusively whith random experiments where all unitary outcomes have equal probability, and where no events are rejected due to not being "representative."

The phrase "representative sample" is found frequently in books and articles, but in the vast majority of cases the term is used incorrectly. Statistics deals with random samples, not with "representative" samples. Excluding some very specific methods of sampling, there is no such thing as a “representiative” sample. Most importantly, the validity of a sample should be based on the method used to obtain the sample and the methods and models used to make estimations and inferences based on the sample. In general, the validity of a sample should not be based on the characteristics of the sample assessed after the sample is obtained. Beware of “representative” samples. All the equations we will use in this book apply to random samples, not representative ones.

It is possible that some authors use the term “representative” to characterize samples that were selected following a proper method and whose elements are processed using the correct formulas specifically derived for the sampling method used. In this case, of course that we want “representative” samples. Here, the term is simply used to mean “properly obtained and analyzed” samples.

[@Pietra1926] Wrote *A particular case of non-representative sampling* to point out that randomly sampling persons and averaging their family sizes is not a “representative” sampling process. Intuitively, one should suspect that something might be amiss when we sample objects type A (individuals) and use the formulas for A to make statements about objects type B (families). By sampling individuals randomly, the probability of a family being selected is proportional to its size. Simply using the average to estimate the mean does not work, because a direct use of the average to estimate the mean assumes equal-probability sampling. Presumably, the author of the paper used the term non-representative to mean “unequal probability” sampling. These days there are multiple sampling methods and corresponding formulas that do not use simple random sampling and do use unequal probability sampling.

Incidentally, and to remark the importance of formulating questions carefully, the sampling mentioned by [@Pietra1926] was not suitable to estimate average family size by simply taking the sample average for the individuals selected by simple random sampling. However, it was correct to take the simple average to estimate the average size of family *experienced by individuals*. To futher understand this issue, think about the average class size in UCD vs. the average class attended by students. How would you go about calculating each one if you had the complete database of courses and enrollments?

### 1.7.3 No Biased samples allowed

"Biased" is another term unfortunately used to describe samples. The best scenario would be to define biased samples as samples obtained using a method that does not assign equal probability to all elements of the population. But there is a better term that does not have the negative connotation: Unequal probability sampling. It is probably best not to use the term "bias" to talk about samples or sampling methods and to reserve it to be used as a property of some estimators.

There are many sampling methods that use unequal probability. If they are used in conjunction with the proper estimators and analysis procedures, there should not be a problem. As a matter of fact, unequal sampling probability may be the only approach in some situations and it is not a problem for as long as the probabilities of any unit being selected is known and greater than 0. Assigning a zero probability to some elements of the population is equivalent to defining a new populations without them. Obviously, the samples will not have information about the original but about the reduced population. Therefor, another interpretation of the phrase "biased sample" would be that the sample does not come from the population about which we want to make statements.

- Samples have to be obtained using a random process in which all units of the population have a strictly positive and known probability of being selected, even if the probabilites differ among units.

Statistics deals with sampling where the probability of any unit being included in the sample is greater than zero (e.g., no units are categorically excluded) and known. This is the only way one can calculate the statistical properties of estimators, particularly their variance and bias.

### 1.7.4 Sampling methods

We get samples to estimate somethig about the whole population. Sampling or survey design methods are particularly important when we are not doing manipulative studies but observational ones. In manipulative experiments we will always assume that treatments are assigned at random to the experimental units, with certain simple modifications or restrictions applied to designs more advanced than those treated in this text. Experimental units should be selected as a proper sample of the population whose response to treatments we want to estimate. Unfortunately, most designed experiments DO NOT use randomly selected experimental units, so their results have to be taken with caution. For example, medical research is based on either patients that seek medical treatment or volunteers. Neither constitutes a random sample of the population. Agricultural experimentation is conducted with animals either maintained or bred for experimentation, or selected haphazardly from what is available (sample of convenience). Therefore, there is a real and potentially large gap between the samples for which theory and formulas were derived, and actual samples used.

**Sampling design** is the procedure used to select the units for the sample from the population, such that each possible sample has a positive and known probability of being selected.

We will consider only designs where all samples have equal probability.

**Random sampling with equal probabilities without replacement** The simplest sampling design is random sampling where all units have equal probability of being selected. If we obtain samples of size r out of a finite population with N elements, there are $C_r^N$ different samples possible without replacement, each with equal probability.

**Random sampling with equal probability with replacement** In this case there are $N^r$ different samples possible, all with equal probability.

When populations are infinite, as in most of the cases we treat in the rest of the book, the difference between sampling with and without replacement disappears because, regardless of the sample size n, the ratio of sample size to population size is infinitesimal.

### 1.7.5 Estimators

Populations are characterized by parameters. Parameters are constant values that are typically unknown. Samples are used to estimate parameters or to make statements about differences in parameters. Although for the Normal distribution the parameters are the same as the mean and the variance, for other distributions this is not the case. The question arises then: how do we estimate parameters based on samples??? There are several methods to estimate parameters, some of which are:

- Least-squares
- Maximum likelihood

- Method of moments

Least-squares estimators are formulas to estimate the parameters that minimize the sum of the squared differences between the observations and the values predicted by using the estimated parameter.

Maximum likelihood estimators yield estimated parameter values that maximize the probability of observing the data actually obtained.

The method of moments derives equations to calculate the parameters as a function of the moments (say mean and variance for example). Then it calculates the sample moments and inserts the sample moment values in the derived equations to get estimated parameters.

The sample average is the least squares, maximum likelihood and method-of-moments estimator for the mean of a normal distribution.

Parameters are estimated by doing calculations with the observed sample values. The results of doing calculations based on sample values are called **statistics** Obviously, this is a different meaning of the term **statistics* in the “science of statistics”.

- A **statistic** is a function of sample values, and therefore, a random variable.
- An **estimator** is a statistic used to estimate a parameter.

For example, the **sample average** is a statistic that is typically used to estimate a parameter: the **mean** of a normal population. However, the mean is not necessarily a parameter of all distributions, so we need to come up with methods to estimate parameters in general, not just the mean for normal distributions. For example, the mean is a moment but not a parameter in the Binomial distribution.

## 1.8 Sampling distributions

Statistic: a function of the values observed in a sample. Estimator: a statistic that is used to estimate a parameter. Estimate: the value of a statistic for a given sample.

Example: a sample of 3 values is drawn from the standard normal distribution. Those values are -0.2359879422, 1.0029272486, -1.0054658377. the sample average or simply “average” is a statistic that consists of averaging all observed values -0.0795088438 and the sample variance is 1.0267749785

## 1.9 Estimation vs. inference and prediction

Recall the use of the binomial or Poisson distributions to calculate probabilities of events. In those cases we used models whose parameters were known to calculate exact probabilities for certain events of interest. For example, we calculated the probability that at least 5 seeds germinate in a sample of 10 seeds. If the germination actually follows the distribution used, then the probability calculated is correct and known with certainty. We could say that before we put a set of seeds to germinate, that we “predit that 5 seeds will germinate with a probability equal to xx.” We make what is called a PREDICTION for the results of a random experiment (germinating the seeds). If we repeat the experiment a large number of times, we should be able to corroborate that the number of times 5 seeds germinate divide by the number of experiments is xx.

**A PREDICTION is a statement about the value that a random variable will take in a (future) random experiment.**

The situation above is very different from a situation in which we do not know the parameter $p$ or probability of success of the binomial distribution. In this case we need to ESTIMATE the parameter in orer to be able to make any predictions. Parameters can be estimated by using data. For example, we could test 1000 seeds and calculate the proportion that germinate. This estimate is called $\hat{p}$. Now this estimate $\hat{p}$ of $p$ is itself a random variable and it has a variance and expected values that can be estimated.

**An ESTIMATION is a statement about the current unknown value of a parameter or function of parameters.**

In traditional frequentist statistics, ESTIMATION is the process of obtaining estimates for parameter values, whereas INFERENCE refers to making probabilistic statements about parameters or functions of parameters.

**TERMINOLOGY WARNING!** Different authors use the same terms to refer to different things. Authors are free to use language in any way they see fit. Unfortunately, this means that the reader has to be adaptive and understand how each author uses the terms. We emphasize the knowledge of the list of concepts and the concepts themselves, without being adamant about our usage.

Leaving names aside, some of the various types of statements we can make are:

1. Predict the outcome of a random experiment or the value of random variable.
2. Estimate the unknown value of a parameter.
3. Make a confidence interval for the parameter.

For statement 1 we need to know the parameters (or their estimates) and the distribution of the random variable. For statement 2 we only need to estimate the parameter using some estimator. For statement 3 we need to have the estimate and the sampling variance of the estimator. An example will make this clearer.

Suppose that we are interested in the square of people's height. This quantitiy is used in the calculation of the Body Mass Index. Further suppose that height follows a normal distribution with mean $\mu$ and variance $\sigma^2$. ****TO BE COMPLETED****

## 1.10 Exercises and Solutions

Most exercises with solutions in the book *Probability and Statistics Applications for Environmental Science* by [@PSAES2007] are relevant for this chapter. The book can be downloaded from a UCD IP address (either in UCD or using the library VPN) from https://www.taylorfrancis.com/books/9781420007824.

### 1.10.1 Exercise

Find processes that could be described with the more common distribution functions, sample them amd create histograms for the results. Determine if the results look like the hypothesized distribution.

### 1.10.2 Exercise

This is a game. Each player starts with 10 cents in pennies. The bank (instructor) chooses with probability p1 to either roll a die or flip a coin 5 times and reports the number of dots or the number of heads + 1. Students bet on whether the bank used the coin or the die. This can be extended to increasing number of “draws” before stoping the bet. Use Bayes rule to calculate the posterior for coin and die.

## 1.11 Homework

1. 
2. Make a frequency table where two random variables are not correlated and are not independent. Use calculations to show both properties are as requested.

## 1.12 Laboratory Exercises

### 1.12.1 Plant Sciences

### 1.12.2 Animal Sciences

---

[1] Amazingly, some distributions do not have mean or variance, for example because the mean or variance involve sum of series that do not converge.↩