Required Math Skills and Symbols

Learning Objectives for Chapter

1. Pair symbols with corresponding concepts or definitions.
2. Express sums using summation symbol and expand sums defined with the summation symbol.
3. Calculate sum of squared deviations for a set of numbers.
4. Explain in your own words what the sum of squares of a set of numbers describes.

5. Identify the symbols that correspond to standard deviation, variance, population mean, and sample average.

6. List the different strategies we can use to find the optimal value for an unknown variable.

Data Columns and Summation

Summations, the sum of several numbers that are organized with subscripts, is used all the time in statistics. In order to be able to read equations and understand the concepts in them, you have to be familiar with summations. The greek letter capital sigma Σ is used to represent summation. For example, $\sum_{i=1}^{4} Y_i$ represents the sum $Y_1 + Y_2 + Y_3 + Y_4$. Sometimes, for simplicity, when we want to sum over all possible values of the subscript i from $i = 1$ to the maximum but unspecified value of $i = k$, we write $\sum_{i=1}^{k} Y_i$, meaning "sum of all values of Y_i."

We can write a symbol for all the Y_i to avoid having to write each one of them. All the values of Y in the correct order constitute a *vector* of values. We use bold letters to refer to vectors. Suppose we have a vector with the numbers 7, 5, 9 and 6. We write this as:

$$\mathbf{Y} = \begin{pmatrix} Y_1 & Y_2 & Y_3 & Y_4 \end{pmatrix} = \begin{pmatrix} 7 & 5 & 9 & 6 \end{pmatrix} = (Y_i) \quad i = 1, \ldots, 4$$

where it is impied that $Y_1 = 7$, $Y_2 = 5$, $Y_3 = 9$ and $Y_4 = 6$. There are no calculations or concepts here other than the fact that vectors are sets of numbers with order, and that we can refer to them with symbols. In R we can define the vector and give it values using the `c` function, which stands for "combine."

```r
Y <- c(7, 5, 9, 6)
print(Y)
```

```
## [1] 7 5 9 6
```

```r
# The "[1]" on the left simply indicates that 7 is the first element in the vector.
```

These summation equations have parallels in R code. In R, we put all values of Y in a vector, and the subscripts are automatically assigned or implicit based on the position of each value of Y. You can think of the vector being a column, so the positions of the vector elements are the row numbers.

We can refer to each element in Y by using its row number or index number inside brackets and right next to the name of the vector. For example, `Y[2]` is the second element of Y, which is the number 5. Using different expressions instead of a single number we can extract any parts of Y we need to operate with.

```r
# Extract the third element of Y
Y[3]
```

```
## [1] 9
```

```r
# The "[1]" on the left simply indicates that 3 is the first element
# in the vector represented by the third element of Y.
# Extract elements 2 to 3 of Y
Y[2:3]
```

```
## [1] 5 9
```

```r
# The "[1]" on the left simply indicates that 5 is the first element
# in the vector represented by the third and forth elements of Y.

# Extract elements 1, 3 and 4 of Y
Y[c(1, 3:4)]
```

```
## [1] 7 9 6
```

Consider the following to get an idea of what the symbols mean and how they work.

$$\sum \mathbf{Y} = \sum_{i=1}^{i=4} Y_i = Y_1 + Y_2 + Y_3 + Y_4 = 7 + 5 + 9 + 6 = 27$$

Because most R operations are vectorized, if we want to use the whole vector we do not need to specify subscripts, R takes `sum(Y)` to mean "sum all elements of Y." To sum parts of a vector we need to use the extraction method.

```r
sum(Y)
```

```
## [1] 27
```

$$\sum_{i=2}^{i=4} Y_i = Y_2 + Y_3 + Y_4 = 5 + 9 + 6 = 20$$

```r
sum(Y[2:4])
```

```
## [1] 20
```

The subscripts for summations can themselves be equations, which makes these formulas very versatile, although in this course we will not be using many complicated subscripts.

$$\sum_{i=1}^{i=3} Y_i = \sum_{i=2}^{i=4} Y_{i-1} = Y_1 + Y_2 + Y_3 = 7 + 5 + 9 = 21$$

$$\begin{aligned}\sum_{i=1}^{i=3} (Y_i + Y_{i+1}) &= (Y_1 + Y_2) + (Y_2 + Y_3) + (Y_3 + Y_4) \\ &= Y_1 + Y_2 + Y_2 + Y_3 + Y_3 + Y_4 \\ &= 7 + 5 + 5 + 9 + 9 + 6 = 41\end{aligned}$$

As usual, there are many ways to achieve a task in R. To perform the sum above we can use the following alternatives:

```r
sum(Y[1:3] + Y[2:4])
```

```
## [1] 41
```

```r
sum(Y[1:3]) + sum(Y[2:4])
```

```
## [1] 41
```

```
i <- 1:3
sum(Y[i] + Y[i + 1])
```

```
## [1] 41
```

In order to explain further properties of the summation, let's use a second vector represented by the letter **X**, where

$$\mathbf{X} = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 \end{pmatrix} = \begin{pmatrix} 3 & 2 & 4 & 5 \end{pmatrix} = (X_i) \quad i = 1, \ldots, 4$$

```
X <- c(3, 2, 4, 5)
print(X)
```

```
## [1] 3 2 4 5
```

The notation is very general and can be used to express sums like the following sum of products of corresponding elements in **X** and **Y**:

$$\begin{aligned} \sum_{i=1}^{i=4} (X_i\; Y_i) &= (X_1\; Y_1) + (X_2\; Y_2) + (X_3\; Y_3) + (X_4\; Y_4) \\ &= 3 \times 7 + 2 \times 5 + 4 \times 9 + 5 \times 6 = 97 \end{aligned}$$

Two-dimensional Data Tables and Summation

The vectors **X** and **Y** used above are 1-dimensional; they are single columns of numbers. Frequently, data is organized in more than one dimension, as in a 2-dimensional table with rows and columns. A conventional and very clear way to represent those tables is a matrix where each element is identified by its "address" in the table given by the row and column numbers, always in that order. The subscript i refers to rows and j refers to columns. We illustrate this with an example in which soil moisture was measured under wheat, corn and tomatoes in each of 7 fields (numbers are fictitious). The matrix is built by using rows for fields and columns for crops. Note that the subscripts are not numbers with 2 digits, but two single-digit integer written together; subscript 62 does **not** mean 6 times 10 plus 2 times 1, but simply the sixh row and second column. For situations when there are more than 9 rows and or columns, we use a separator between the row and column numbers to avoid ambiguity (for example, to be able to tell if Y_{162} is row 1 column 62 or row 16 column 2 we use $Y_{1,62}$ or $Y_{16,2}$. The comma is ommitted when each subscript is never greater than 9.

$$\mathbf{Y} = \begin{pmatrix} Y_{1,1} & Y_{1,2} & Y_{1,3} \\ Y_{2,1} & Y_{2,2} & Y_{2,3} \\ Y_{3,1} & Y_{3,2} & Y_{3,3} \\ Y_{4,1} & Y_{4,2} & Y_{4,3} \\ Y_{5,1} & Y_{5,2} & Y_{5,3} \\ Y_{6,1} & Y_{6,2} & Y_{6,3} \\ Y_{7,1} & Y_{7,2} & Y_{7,3} \end{pmatrix} = \begin{pmatrix} 34 & 30 & 29 \\ 32 & 26 & 27 \\ 27 & 26 & 24 \\ 37 & 33 & 34 \\ 25 & 24 & 23 \\ 23 & 22 & 20 \\ 30 & 27 & 26 \end{pmatrix}$$

In R we can define a matrix or a dataframe with rows and columns. We use a dataframe because it is the most common form for data in R and in science in general. Instead of Y, in the code below we call the data "Ydata" to avoid conflicts with the names of the vectors used above. First we make vectors of data for each crop and then we join them into a data frame with the `cbind` function, which stands for "column bind."

```r
wheat <- c(34, 32, 27, 37, 25, 23, 30)
corn <- c(30, 26, 26, 33, 24, 22, 27)
tomat <- c(29, 27, 24, 34, 23, 20, 26)

Ydata <- as.data.frame(cbind(wheat, corn, tomat))

Ydata
```

```
##   wheat corn tomat
## 1    34   30    29
## 2    32   26    27
## 3    27   26    24
## 4    37   33    34
## 5    25   24    23
## 6    23   22    20
## 7    30   27    26
```

```r
Ydata[5, 2] # Soil moisture for field 5 with corn
```

```
## [1] 24
```

In general, for a table of values with k rows and r columns we write:

$$\mathbf{Y}_{k,r} = \begin{pmatrix} Y_{1,1} & Y_{1,2} & \cdots & Y_{1,r} \\ Y_{2,1} & Y_{2,2} & \cdots & Y_{2,r} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{k,1} & Y_{k,2} & \cdots & Y_{k,r} \end{pmatrix}$$

If we want to average all values in the table we need to use a double summation to indicate that the sum is supposed to proceed over rows and columns. As an exercise to make sure you understand the subscripts, write down the matrix **Y** with all subscripts without looking at the book

$$\sum_{i=1}^{i=7} \sum_{j=1}^{j=3} \frac{Y_{ij}}{7 \times 3} = \frac{1}{7 \times 3} \sum_{i=1}^{i=7} \sum_{j=1}^{j=3} Y_{ij}$$

$$= \frac{1}{7 \times 3} \left(\sum_{j=1}^{j=3} Y_{1j} + \sum_{j=1}^{j=3} Y_{2j} + \sum_{j=1}^{j=3} Y_{3j} + \sum_{j=1}^{j=3} Y_{4j} + \sum_{j=1}^{j=3} Y_{5j} + \sum_{j=1}^{j=3} Y_{6j} + \sum_{j=1}^{j=3} Y_{7j} \right)$$

$$\begin{aligned} = \frac{1}{7 \times 3} & (Y_{11} + Y_{12} + Y_{13} + Y_{21} + Y_{22} + Y_{23} + Y_{31} + Y_{32} + Y_{33} \\ & + Y_{41} + Y_{42} + Y_{43} + Y_{51} + Y_{52} + Y_{53} + Y_{61} + Y_{62} + Y_{63} + Y_{71} + Y_{72} + Y_{73}) \end{aligned}$$

$$\begin{aligned} & = \frac{1}{7 \times 3} (34 + 30 + 29 + 32 + 26 + 27 + 27 + 26 + 24 + 37 + 33 + 34 \\ & + 25 + 24 + 23 + 23 + 22 + 20 + 30 + 27 + 26) = \frac{579}{21} = 27.57 \end{aligned}$$

Using R we can calculate the sums and the average very easily. Note that the functions `nrow` and `ncol` extract the number of rows and columns of a data frame or matrix. The function `mean` give the average of the table.

```r
sum(Ydata) # Works only when all columns are numeric.
```

```
## [1] 579
```

```r
nrow(Ydata) # Ydata has 7 rows
```

```
## [1] 7
```

```r
ncol(Ydata) # and 3 columns.
```

```
## [1] 3
```

```r
sum(Ydata) / (nrow(Ydata) * ncol(Ydata))
```

```
## [1] 27.57143
```

```r
mean(as.matrix(Ydata)) # Works only when all columns are numeric.
```

```
## [1] 27.57143
```

If we just want the average for corn we specify just the column for corn (column 2) and sum over all rows for column 2. If we want the results for a field or a set of fields, we specify the corresponding row numbers. Let's calculate the average moisture for corn and then the average for fields 3, 4, 5 and 7.

$$corn\ average = \frac{1}{7}\left(\sum_{i=1}^{i=7} Y_{i2}\right)$$

$$= \frac{1}{7}(Y_{12} + Y_{22} + Y_{32} + Y_{42} + Y_{52} + Y_{62} + Y_{72})$$

$$= \frac{1}{7}(30 + 26 + 26 + 33 + 24 + 22 + 27) = 26.86$$

$$fields\ 3,\ 4,\ 5\ \&\ 7\ average = \frac{1}{4 \times 3}\left(\sum_{j=1}^{j=3} Y_{3j} + \sum_{j=1}^{j=3} Y_{4j} + \sum_{j=1}^{j=3} Y_{5j} + \sum_{j=1}^{j=3} Y_{7j}\right)$$

$$= \frac{1}{4 \times 3}(Y_{31} + Y_{32} + Y_{33} + Y_{41} + Y_{42} + Y_{43} + Y_{51} + Y_{52} + Y_{53} + Y_{71} + Y_{72} + Y_{73})$$

$$= \frac{1}{4 \times 3}(27 + 26 + 24 + 37 + 33 + 34 + 25 + 24 + 23 + 30 + 27 + 26)$$

$$= \frac{336}{12} = 28.0$$

We obtain those averges in R with the following code. Note that when we leave the place for row (or column) number empty, it is interpreted as all rows (or columns).

```r
mean(Ydata[, "corn"]) # Average for corn across fields
```

```
## [1] 26.85714
```

```r
sum(Ydata[, 2])
```

```
## [1] 188
```

```r
length(Ydata[, 2])
```

```
## [1] 7
```

```r
sum(Ydata[, 2]) / length(Ydata[, 2])
```

```
## [1] 26.85714
```

Models

In statistics we use models all the time, and it is necessary to make those models explicit. Explicit models make it easier to understand and critically evaluate the use of statistics. Specifically, we can determine what parts of the model may be inadequate to solve the problem at hand, and we can modify the model to improve it.

Say that we are interested in studying the milk production per dairy cow in $kg\ day^{-1}$ for all dairy cows in the US today. For example, we want to estimate what proportion of cows produce less than 15 $kg\ day^{-1}$. One way to do this is to create a *model* for the statistical distribution of milk production, take a sample to estimate the parameters of the model and then use the model with estimated parameters to make the estimation of the proportion of cows for which production is less than 15 $kg\ day^{-1}$ or any other quantity.

A first attempt to estimate the needed proportion is to *model* milk production per cow as a normall distributed random variable with unknown mean and variance. Using the letter Y to represent the milk production of a cow we write $Y \sim N(\mu, \sigma^2)$, which is read "Y is a random variable with normal distribution, with mean μ and variance σ^2." Sometimes we write σ, the standard deviation, instead of σ^2. Obviously, the standard deviation is the square root of the variance.

The model is not correct, and it could not be correct, for a number of reasons. First, normal distributions can take any values between $-\infty$ and $+\infty$, whereas milk production is zero or positive, but it cannot be negative. Moreover, if we are considering the population to be all the dairy cows in the US, althogh there are many dairy cows in the US (USDA estimated 9.3 million cows and heifers), the number is finite, whereas the normal dsitribution is for infinite populations. Yet, these flaws of our model are not important.

> Models are not supposed to be perfect representations of reality, they are supposed to be **useful** representations of reality.

Assuming that the variance of milk production is small relative to the mean, the fact that production has to be positive is not a problem, because a tiny and irrelevant piece of the distribution is expected to be below zero. The fact the the numebr of cows is finite is not a problem because we can either think of the total population as a very, very large sample of a truly infinite theoretical cow population or simply use the normal as an approximation to the truly discrete and finite real population.

The simple model that we are using can also be expressed as follows:

$$Y_i = \mu + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

where ϵ is a random variable with a normal distriution with mean 0 and variance σ^2. This is exactly the same model as before, but now we have isolated the deviations of each cow from the mean for the whole population which we call "errors" not because there is anthing wrong, but because we choose not to be interested in why cows differ in milk production per day.

In order to use this model we can get a random sample of r cows from all the values of milk production and based on that sample we can estimate the mean as

$$\hat{\mu} = \frac{1}{r}\sum_{i=1}^{i=r} Y_i$$

and the variance as

$$\hat{\sigma^2} = \frac{1}{r-1}\sum_{i=1}^{i=r}(Y_i - \hat{\mu})^2$$

where $e = \hat{\epsilon} = (Y_i - \hat{\mu})$ are the deviations of each cow from the estimated overall mean. These equations are presented with more detail in a later chapter (insert cross reference to chapter 06). Finally, the proportion of cows whose daily milk production is less than 15 $kg\ day^{-1}$ is estimated as the area under the normal distribution curve between $-\infty$ and 15 $kg\ day^{-1}$. Using the capital letter$P$ to indicate “probability” we can write:

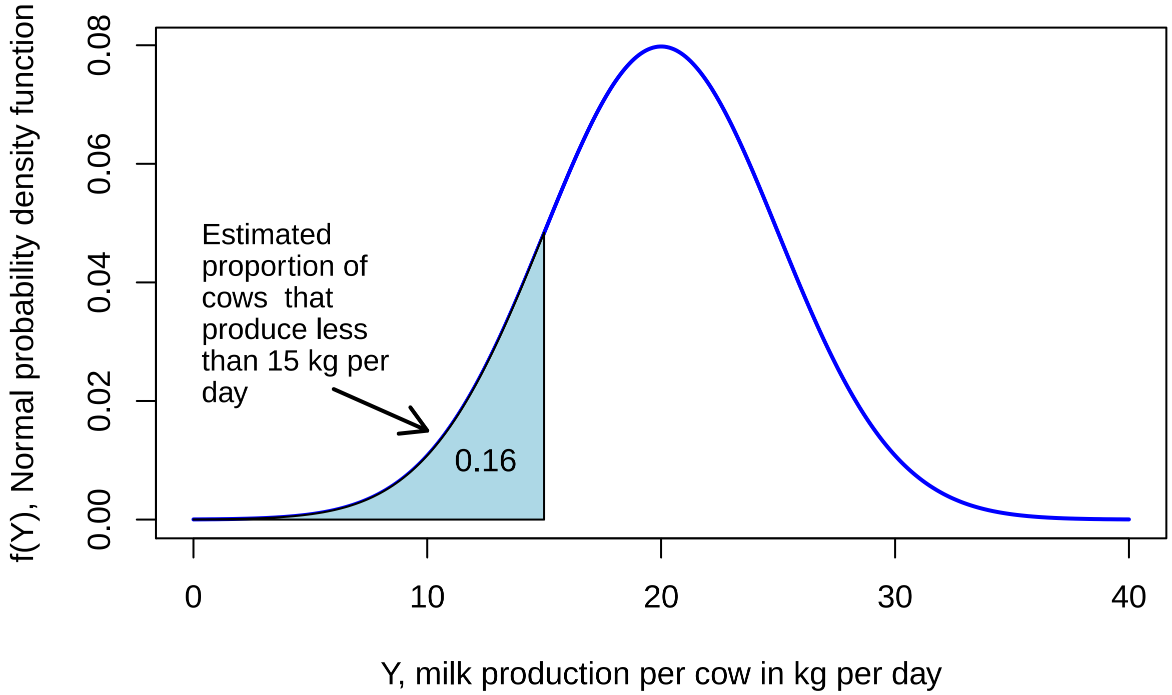
$$\hat{P}(Y \leq 15\ kg\ day^{-1}) = area\ under\ curve = \int_{-\infty}^{15} f(Y)\ \mathrm{d}Y$$

$$where\ f(Y) = \frac{1}{\hat{\sigma}\sqrt{2\pi}}\ \exp\left(-\left(x-\hat{\mu}\right)^2/2\hat{\sigma}^2\right)$$

Suppose that the estimated mean and variance were 20 $kg\ day^{-1}$ and 25 $kg^2\ day^{-2}$, then the area ander the curve would be calculated in R as follows (note how the code is written in several lines and indented for better readability):

```r
pnorm(           # function that calculates areas unde the normal
   q = 15,       # quantile desired, max. milk production in set
   mean = 20,    # mean of the normal distribution
   sd = 5        # standard distribution of the normal distribution
)
```

```
## [1] 0.1586553
```



Deviations, sums of squares, sums of absolute deviations

The difference between an observation and the mean is called a *deviation*, *residual* or *error*. The specific name depends on the context, but in general, we will use *deviation* for the total difference between observation and overall average, *residual* for the difference between an observation and the value estimated by a model, and *error* for the random component of a model.

Geometry of sums of squares

Remember the Pythagorean theorem? In a right triangle the square of the hypothenuse equals the sum of the squares of the sides. This can be extended to a polyhedron with right angles everywhere. The square of a diagonal in 3D is the sum of the three sides of the polyhedron. This square is also the sum of the square of the bottom hypothenuse and the vertical side squared. ** Add figure with labels** Imagine that we have a sample with 3 observations. Make a box where the lenght of each side is one of the observations and then consider the 3D diagonal. The sum of squares of the deviations is equal to the length of the diagonal.

The total sum of squares …

Optimization

One way to "guess" or estimate the mean of a distribution that can be sampled directly is to take the average. The average minimizes the sum of the squared deviations of the obervations. Minimization of sums of squares is the most common method we will use to make estimates, but it is not the onlly method. Other method is to minimize the sum of the absolute deviations. The median of a set of observations is the value that minimizes the absolute deviations. "Absolute" means that all deviations are made positive.

Show an interactive graph where the values SS and SAD are shown as functions of the guessed mean. With and without outliers.

In many cases, the optimal value can be found analytically, which means that we can use an equation that yields the best estimate directly. This formula is obtained by taking the first derivative of the equation that calculates the SS as a function of the estimate, set it to zero and solve for the estimate. For more advanced statistical methods, there are no equations available, so numerical computations are used to approximated the optimal estimates.

Show pseudocode for an algorithm to find out what number I am thinking of or to numerically approximate a root of a polynomial

Say what a polynomial is.

Complete list and definition of symbols used in the book

Exercises and Solutions

Homework

Laboratory Exercises

Plant Sciences

Animal Sciences