# 008 - Probability

EPIB 607

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca
https://sahirbhatnagar.com/EPIB607/ChapProbability.html

slides compiled on September 19, 2021

## Introduction to probability

## Variability of a population

## Variability of a random variable

# Introduction

- This section extends the notion of variability that was introduced in the context of data to other situations.
- The variability of the entire **population** and the concept of a **random variable** is discussed.
- These concepts are central for the development and interpretation of statistical inference.

# Objectives

- Consider the distribution of a variable in a population and compute parameters of this distribution, such as the mean and the standard deviation.

# Objectives

- Consider the distribution of a variable in a population and compute parameters of this distribution, such as the mean and the standard deviation.
- Become familiar with the concept of a random variable.

# Objectives

- Consider the distribution of a variable in a population and compute parameters of this distribution, such as the mean and the standard deviation.
- Become familiar with the concept of a random variable.
- Understand the relation between the distribution of the population and the distribution of a random variable produced by sampling a random subject from the population.

# Objectives

- Consider the distribution of a variable in a population and compute parameters of this distribution, such as the mean and the standard deviation.
- Become familiar with the concept of a random variable.
- Understand the relation between the distribution of the population and the distribution of a random variable produced by sampling a random subject from the population.
- Identify the distribution of the random variable in simple settings and compute its expectation and variance.

# Variability in Data

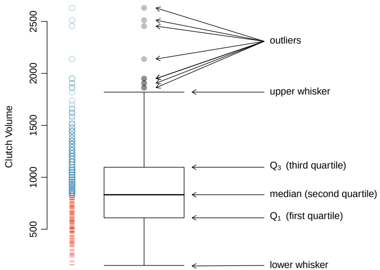



Figure: A boxplot and dot plot of `clutch.volume` in the `frog` dataset from the `oibiostat` package. The horizontal dashes indicate the bottom 50% of the data and the open circles represent the top 50%.

- We previously examined the variability in data.
- In the statistical context, data is obtained by selecting a sample from the target population and measuring the quantities of interest for the subjects that belong to the sample.
- Different subjects in the sample may obtain different values for the measurement, leading to variability in the data.
- Numerical summaries may be computed in order to characterize the main features of the variability, e.g., mean, median, sample variance, sample standard deviation, inter-quartile range

# Two other forms of variability

- The subject of this section is to introduce two other forms of variability, variability that is not associated, at least not directly, with the data that we observe.

# Two other forms of variability

- The subject of this section is to introduce two other forms of variability, variability that is not associated, at least not directly, with the data that we observe.
    1. Population variability
    2. Variability of a random variable

# Two other forms of variability

- The subject of this section is to introduce two other forms of variability, variability that is not associated, at least not directly, with the data that we observe.
    1. Population variability
    2. Variability of a random variable
- The notions of variability that will be presented are abstract, they are not given in terms of the data that we observe, and they have a mathematical-theoretical flavor to them.

# Two other forms of variability

- The subject of this section is to introduce two other forms of variability, variability that is not associated, at least not directly, with the data that we observe.
    1. Population variability
    2. Variability of a random variable
- The notions of variability that will be presented are abstract, they are not given in terms of the data that we observe, and they have a mathematical-theoretical flavor to them.
- At first, these abstract notions may look to you as a waste of your time and may seem to be unrelated to the subject matter of the course.

# Two other forms of variability

- The subject of this section is to introduce two other forms of variability, variability that is not associated, at least not directly, with the data that we observe.
    1. Population variability
    2. Variability of a random variable
- The notions of variability that will be presented are abstract, they are not given in terms of the data that we observe, and they have a mathematical-theoretical flavor to them.
- At first, these abstract notions may look to you as a waste of your time and may seem to be unrelated to the subject matter of the course.
- The opposite is true. The very core of statistical thinking is relating observed data to theoretical and abstract models of a phenomena.

# Two other forms of variability

- The subject of this section is to introduce two other forms of variability, variability that is not associated, at least not directly, with the data that we observe.
    1. Population variability
    2. Variability of a random variable
- The notions of variability that will be presented are abstract, they are not given in terms of the data that we observe, and they have a mathematical-theoretical flavor to them.
- At first, these abstract notions may look to you as a waste of your time and may seem to be unrelated to the subject matter of the course.
- The opposite is true. The very core of statistical thinking is relating observed data to theoretical and abstract models of a phenomena.
- The abstract notions of variability that are introduced in this chapter, and are extended in the subsequent chapters, are the essential foundations for the practice of statistics

# Population variability

- We will examine the data set `heights_sample.csv` available at https://github.com/sahirbhatnagar/EPIB607/blob/master/inst/data/heights_sample.csv which contains data on the sex and height of a sample of 100 observations.

# Population variability

- We will examine the data set `heights_sample.csv` available at https://github.com/sahirbhatnagar/EPIB607/blob/master/inst/data/heights_sample.csv which contains data on the sex and height of a sample of 100 observations.
- We will also consider the sex and height of all the members of the population from which the sample was selected available at https://github.com/sahirbhatnagar/EPIB607/blob/master/inst/data/heights_population.csv

# Population variability

- We will examine the data set `heights_sample.csv` available at https://github.com/sahirbhatnagar/EPIB607/blob/master/inst/data/heights_sample.csv which contains data on the sex and height of a sample of 100 observations.
- We will also consider the sex and height of all the members of the population from which the sample was selected available at https://github.com/sahirbhatnagar/EPIB607/blob/master/inst/data/heights_population.csv
- The size of the relevant population is 100,000, including the 100 subjects that composed the sample.

# Population variability

- We will examine the data set `heights_sample.csv` available at https://github.com/sahirbhatnagar/EPIB607/blob/master/inst/data/heights_sample.csv which contains data on the sex and height of a sample of 100 observations.
- We will also consider the sex and height of all the members of the population from which the sample was selected available at https://github.com/sahirbhatnagar/EPIB607/blob/master/inst/data/heights_population.csv
- The size of the relevant population is 100,000, including the 100 subjects that composed the sample.
- When we examine the values of the height across the entire population we can see that different people may have different heights. This variability of the heights is the **population variability**

# Variability of a random variable

- The other abstract type of variability, the **variability of a random variable**, is a mathematical concept.

# Variability of a random variable

- The other abstract type of variability, the **variability of a random variable**, is a mathematical concept.
- The aim of this concept is to model the notion of randomness in measurements or the uncertainty regarding the outcome of a measurement.

# Variability of a random variable

- The other abstract type of variability, the **variability of a random variable**, is a mathematical concept.
- The aim of this concept is to model the notion of randomness in measurements or the uncertainty regarding the outcome of a measurement.
- In particular we will initially consider the variability of a random variable in the context of selecting one subject at random from the population.

# Variability of a random variable

- The other abstract type of variability, the **variability of a random variable**, is a mathematical concept.
- The aim of this concept is to model the notion of randomness in measurements or the uncertainty regarding the outcome of a measurement.
- In particular we will initially consider the variability of a random variable in the context of selecting one subject at random from the population.
- Random variables provide models for randomness and uncertainty in measurements.

# Variability of a random variable

- The other abstract type of variability, the **variability of a random variable**, is a mathematical concept.
- The aim of this concept is to model the notion of randomness in measurements or the uncertainty regarding the outcome of a measurement.
- In particular we will initially consider the variability of a random variable in the context of selecting one subject at random from the population.
- Random variables provide models for randomness and uncertainty in measurements.
- Simple examples of such abstract random variables will be provided in this chapter. More examples will be introduced in the subsequent chapters.

# A point to remember

- The variability of the data relates to a concrete list of data values that is presented to us.

# A point to remember

- The variability of the data relates to a concrete list of data values that is presented to us.
- The other types of variability are not associated with quantities we actually get to observe.

# A point to remember

- The variability of the data relates to a concrete list of data values that is presented to us.
- The other types of variability are not associated with quantities we actually get to observe.
- The data for the sample we get to see but not the data for the rest of the population.

# A point to remember

- The variability of the data relates to a concrete list of data values that is presented to us.
- The other types of variability are not associated with quantities we actually get to observe.
- The data for the sample we get to see but not the data for the rest of the population.
- Yet, we can still discuss the variability of a population that is out there, even though we do not observe the list of measurements for the entire population.

# A point to remember

- The variability of the data relates to a concrete list of data values that is presented to us.
- The other types of variability are not associated with quantities we actually get to observe.
- The data for the sample we get to see but not the data for the rest of the population.
- Yet, we can still discuss the variability of a population that is out there, even though we do not observe the list of measurements for the entire population.
- The discussion of the variability in this context is theoretical in its nature. Still, this theoretical discussion is instrumental for understanding statistics.

# Sample of heights

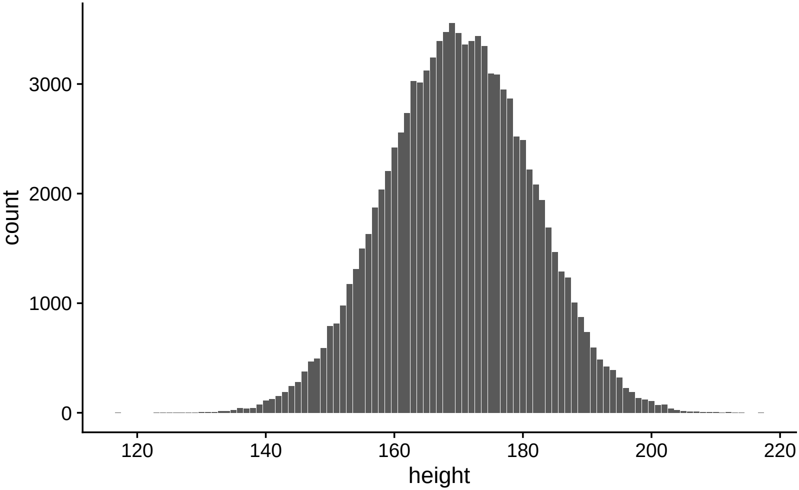
```r
library(dplyr)
library(rio)

heights_sample <- rio::import(
                  here::here("inst/data/heights_sample.csv"))
heights_sample <- heights_sample %>%
                   dplyr::mutate(sex = factor(sex))

summary(heights_sample)

##        id                sex          height
##  Min.   :1538611   FEMALE:54   Min.   :117.0
##  1st Qu.:3339583   MALE  :46   1st Qu.:158.0
##  Median :5105620               Median :171.0
##  Mean   :5412367               Mean   :170.1
##  3rd Qu.:7622236               3rd Qu.:180.2
##  Max.   :9878130               Max.   :208.0

heights_sample %>%
   dplyr::glimpse()

## Rows: 100
## Columns: 3
## $ id     <int> 5696379, 3019088, 2038883, 1920587, 6006813, 4055945, 9263269, ~
## $ sex    <fct> FEMALE, MALE, MALE, FEMALE, MALE, FEMALE, FEMALE, FEMALE, MALE,~
## $ height <int> 182, 168, 172, 154, 174, 176, 193, 156, 157, 186, 143, 182, 194~
```

# Population of heights

```r
heights_population <- rio::import(
                 here::here("inst/data/heights_population.csv"))
heights_population <- heights_population %>%
                    dplyr::mutate(sex = factor(sex))

summary(heights_population)

##        id                sex             height     
##  Min.   :1000082   FEMALE:48888   Min.   :117  
##  1st Qu.:3254220   MALE  :51112   1st Qu.:162  
##  Median :5502618                  Median :170  
##  Mean   :5502428                  Mean   :170  
##  3rd Qu.:7757518                  3rd Qu.:178  
##  Max.   :9999937                  Max.   :217  

heights_population %>%
   dplyr::glimpse()

## Rows: 100,000
## Columns: 3
## $ id     <int> 5696379, 3019088, 2038883, 1920587, 6006813, 4055945, 9263269, ~
## $ sex    <fct> FEMALE, MALE, MALE, FEMALE, MALE, FEMALE, FEMALE, FEMALE, MALE,~
## $ height <int> 182, 168, 172, 154, 174, 176, 193, 156, 157, 186, 143, 182, 194~
```

# Distribution of population of heights

```
library(ggplot2); library(cowplot)
ggplot2::theme_set(cowplot::theme_cowplot())
p <- ggplot(data = heights_population,
            mapping = aes(x = height))
p + geom_bar()
```



# Population mean $\mu$

- The formula of the population mean is:

$$\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}} = \frac{\sum_{i=1}^{N} y_i}{N} .$$

# Population mean $\mu$

- The formula of the population mean is:

$$\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}} = \frac{\sum_{i=1}^{N} y_i}{N} .$$

- Observe the similarity between the definition of the mean for the data and the definition of the mean for the population. In both cases the arithmetic average is computed. The only difference is that in the case of the mean of the data the computation is with respect to the values that appear in the sample whereas for the population all the values in the population participate in the computation.

# Population mean $\mu$

- The formula of the population mean is:

$$\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}} = \frac{\sum_{i=1}^{N} y_i}{N} \; .$$

- Observe the similarity between the definition of the mean for the data and the definition of the mean for the population. In both cases the arithmetic average is computed. The only difference is that in the case of the mean of the data the computation is with respect to the values that appear in the sample whereas for the population all the values in the population participate in the computation.
- In actual life, we will not have all the values of a variable in the entire population. Hence, we will not be able to compute the actual value of the population mean.

# Population mean $\mu$

- The formula of the population mean is:

$$\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}} = \frac{\sum_{i=1}^{N} y_i}{N} \; .$$

- Observe the similarity between the definition of the mean for the data and the definition of the mean for the population. In both cases the arithmetic average is computed. The only difference is that in the case of the mean of the data the computation is with respect to the values that appear in the sample whereas for the population all the values in the population participate in the computation.
- In actual life, we will not have all the values of a variable in the entire population. Hence, we will not be able to compute the actual value of the population mean.
- It's still meaningful to talk about the population mean because this number exists, even though we do not know what its value is. Statistics is about trying to estimate this unknown quantity on the basis of the data we do have in the sample.

# Population variance $\sigma^2$

- The formula of the population variance is defined in a similar way:

$$
\begin{aligned}
\sigma^2 &= \text{The average square deviation in the population} \\
&= \frac{\text{Sum of the squares of the deviations in the population}}{\text{Number of values in the population}} \\
&= \frac{\sum_{i=1}^{N} (y_i - \mu)^2}{N} \ .
\end{aligned}
$$

# Population variance $\sigma^2$

- The formula of the population variance is defined in a similar way:

$$\begin{aligned}\sigma^2 &= \text{The average square deviation in the population} \\ &= \frac{\text{Sum of the squares of the deviations in the population}}{\text{Number of values in the population}} \\ &= \frac{\sum_{i=1}^{N} (y_i - \mu)^2}{N} .\end{aligned}$$

- In R:

```
N <- nrow(heights_population)
var(heights_population$height) * (N - 1) / N
```

```
## [1] 126.1576
```

# Population variance $\sigma^2$

- The formula of the population variance is defined in a similar way:

$$\begin{aligned}\sigma^2 &= \text{The average square deviation in the population}\\ &= \frac{\text{Sum of the squares of the deviations in the population}}{\text{Number of values in the population}}\\ &= \frac{\sum_{i=1}^{N}(y_i-\mu)^2}{N}\ .\end{aligned}$$

- In R:

```
N <- nrow(heights_population)
var(heights_population$height) * (N - 1) / N
## [1] 126.1576
```

- The standard deviation of the population, yet another parameter, is denoted by $\sigma$ and is equal to the square root of the variance.

# Population variance $\sigma^2$

- The formula of the population variance is defined in a similar way:

$$\begin{aligned}
\sigma^2 &= \text{The average square deviation in the population} \\
&= \frac{\text{Sum of the squares of the deviations in the population}}{\text{Number of values in the population}} \\
&= \frac{\sum_{i=1}^{N}(y_i - \mu)^2}{N} \; .
\end{aligned}$$

- In R:

```
N <- nrow(heights_population)
var(heights_population$height) * (N - 1) / N
## [1] 126.1576
```

- The standard deviation of the population, yet another parameter, is denoted by $\sigma$ and is equal to the square root of the variance.
- The typical situation is that we do not know what the actual value of $\sigma$. Yet, we may refer to it as a quantity and we may try to estimate its value based on the data we do have from the sample.

Introduction to probability

Variability of a population

Variability of a random variable

# Variability of a random variable

- As an example, consider taking a sample of size $n = 1$ from the population (a single person) and measuring his/her height. We will apply the function `dplyr::sample_n` to sample one row from the data frame:

```
heights_population %>%
   dplyr::sample_n(size = 1)
```

```
##        id  sex height
## 1 2513465 MALE    170
```

# Variability of a random variable

- As an example, consider taking a sample of size $n = 1$ from the population (a single person) and measuring his/her height. We will apply the function `dplyr::sample_n` to sample one row from the data frame:

```
heights_population %>%
   dplyr::sample_n(size = 1)
```

```
##        id  sex height
## 1 2513465 MALE    170
```

- Let us run the function again:

```
heights_population %>%
dplyr::sample_n(size = 1)
```

```
##        id  sex height
## 1 5178027 MALE    159
```

# Variability of a random variable

- As an example, consider taking a sample of size $n = 1$ from the population (a single person) and measuring his/her height. We will apply the function `dplyr::sample_n` to sample one row from the data frame:

```
heights_population %>%
   dplyr::sample_n(size = 1)
```

```
##        id  sex height
## 1 2513465 MALE    170
```

- Let us run the function again:

```
heights_population %>%
dplyr::sample_n(size = 1)
```

```
##        id  sex height
## 1 5178027 MALE    159
```

- And again

```
heights_population %>%
dplyr::sample_n(size = 1)
```

```
##        id  sex height
## 1 1006901 MALE    174
```

# Random variable

- A random variable is the future outcome of a measurement, **before** the measurement is taken. It does not have a specific value, but rather a collection of potential values with a distribution over these values.

# Random variable

- A random variable is the future outcome of a measurement, **before** the measurement is taken. It does not have a specific value, but rather a collection of potential values with a distribution over these values.
- After the measurement is taken and the specific value is revealed then the random variable ceases to be a random variable! Instead, it becomes data.

# Random variable

- A random variable is the future outcome of a measurement, **before** the measurement is taken. It does not have a specific value, but rather a collection of potential values with a distribution over these values.
- After the measurement is taken and the specific value is revealed then the random variable ceases to be a random variable! Instead, it becomes data.
- One is not able to say what the outcome of a random variable will turn out to be.

# Random variable

- A random variable is the future outcome of a measurement, **before** the measurement is taken. It does not have a specific value, but rather a collection of potential values with a distribution over these values.
- After the measurement is taken and the specific value is revealed then the random variable ceases to be a random variable! Instead, it becomes data.
- One is not able to say what the outcome of a random variable will turn out to be.
- However, one may identify patterns in this potential outcome. For example, knowing that the distribution of heights in the population ranges between 117 and 217 centimeter one may say in advance that the outcome of the measurement must also be in that interval.

# Random variable

- A random variable is the future outcome of a measurement, **before** the measurement is taken. It does not have a specific value, but rather a collection of potential values with a distribution over these values.
- After the measurement is taken and the specific value is revealed then the random variable ceases to be a random variable! Instead, it becomes data.
- One is not able to say what the outcome of a random variable will turn out to be.
- However, one may identify patterns in this potential outcome. For example, knowing that the distribution of heights in the population ranges between 117 and 217 centimeter one may say in advance that the outcome of the measurement must also be in that interval.
- Since there is a total of 3,476 subjects with height equal to 168 centimeters and since the likelihood of each subject to be selected is equal then the likelihood of selecting a subject of this height is 3,476/100,000 = 0.03476. In the context of random variables we call this likelihood probability.

# Random variable

- A random variable is the future outcome of a measurement, **before** the measurement is taken. It does not have a specific value, but rather a collection of potential values with a distribution over these values.
- After the measurement is taken and the specific value is revealed then the random variable ceases to be a random variable! Instead, it becomes data.
- One is not able to say what the outcome of a random variable will turn out to be.
- However, one may identify patterns in this potential outcome. For example, knowing that the distribution of heights in the population ranges between 117 and 217 centimeter one may say in advance that the outcome of the measurement must also be in that interval.
- Since there is a total of 3,476 subjects with height equal to 168 centimeters and since the likelihood of each subject to be selected is equal then the likelihood of selecting a subject of this height is 3,476/100,000 = 0.03476. In the context of random variables we call this likelihood probability.
- In the same vain, the frequency of subjects with hight 192 centimeter is 488, and therefore the probability of measuring such a height is 0.00488. The frequency of subjects with height 200 centimeter or above is 393, hence the probability of obtaining a measurement in the range between 200 and 217 centimeter is 0.00393.

# Sample Space and Distribution

- A random variable refer to numerical values, typically the outcome of an observation, a measurement, or a function thereof.

# Sample Space and Distribution

- A random variable refer to numerical values, typically the outcome of an observation, a measurement, or a function thereof.
- A random variable is characterized via the collection of potential values it may obtain, known as the **sample space** and the likelihood of obtaining each of the values in the sample space (namely, the probability of the value).

# Sample Space and Distribution

- A random variable refer to numerical values, typically the outcome of an observation, a measurement, or a function thereof.
- A random variable is characterized via the collection of potential values it may obtain, known as the **sample space** and the likelihood of obtaining each of the values in the sample space (namely, the probability of the value).
- The probability of each value is the height of the bar above the value, divided by the total frequency of 100,000 (namely, the relative frequency in the population).

# Sample Space and Distribution

- A random variable refer to numerical values, typically the outcome of an observation, a measurement, or a function thereof.
- A random variable is characterized via the collection of potential values it may obtain, known as the **sample space** and the likelihood of obtaining each of the values in the sample space (namely, the probability of the value).
- The probability of each value is the height of the bar above the value, divided by the total frequency of 100,000 (namely, the relative frequency in the population).
- We will denote random variables with capital Latin letters such as $X$, $Y$, and $Z$. Values they may obtain will be marked by small Latin letters such as $x$, $y$, $z$. For the probability of values we will use the letter "P". Hence, if we denote by $Y$ the measurement of height of a random individual that is sampled from the given population then:

$$\mathrm{P}(Y = 168) = 0.03476$$

and

$$\mathrm{P}(Y \geq 200) = 0.00393 \; .$$

# Sample Space and Distribution

- Consider, as yet another example, the probability that the height of a random person sampled from the population differs from 170 centimeter by no more than 10 centimeters. (In other words, that the height is between 160 and 180 centimeters.) Denote by $Y$ the height of that random person. We are interested in the probability $\mathrm{P}(|Y - 170| \leq 10)$

# Sample Space and Distribution

- Consider, as yet another example, the probability that the height of a random person sampled from the population differs from 170 centimeter by no more than 10 centimeters. (In other words, that the height is between 160 and 180 centimeters.) Denote by $Y$ the height of that random person. We are interested in the probability $\mathrm{P}(|Y - 170| \leq 10)$
- In R:

```
Y <- heights_population$height

mean(abs(Y-170) <= 10)

## [1] 0.64541
```

# A note on the previous calculation

- Let us produce a small example that will help us explain the computation of the probability. We start by forming a sequence with 10 numbers:

```
Y <- c(6.3, 6.9, 6.6, 3.4, 5.5, 4.3, 6.5, 4.7, 6.1, 5.3)
```

  The goal is to compute the proportion of numbers that are in the range $[4, 6]$ (or, equivalently, $\{|Y - 5| \leq 1\}$).

# A note on the previous calculation

- Let us produce a small example that will help us explain the computation of the probability. We start by forming a sequence with 10 numbers:

  ```
  Y <- c(6.3, 6.9, 6.6, 3.4, 5.5, 4.3, 6.5, 4.7, 6.1, 5.3)
  ```

  The goal is to compute the proportion of numbers that are in the range $[4, 6]$ (or, equivalently, $\{|Y - 5| \leq 1\}$).

- ```
  abs(Y-5)

  ##  [1] 1.3 1.9 1.6 1.6 0.5 0.7 1.5 0.3 1.1 0.3
  ```
  ```

# A note on the previous calculation

- Let us produce a small example that will help us explain the computation of the probability. We start by forming a sequence with 10 numbers:

```
Y <- c(6.3, 6.9, 6.6, 3.4, 5.5, 4.3, 6.5, 4.7, 6.1, 5.3)
```

The goal is to compute the proportion of numbers that are in the range $[4, 6]$ (or, equivalently, $\{|Y - 5| \leq 1\}$).

- ```
  abs(Y-5)

  ##  [1] 1.3 1.9 1.6 1.6 0.5 0.7 1.5 0.3 1.1 0.3
  ```

- ```
  abs(Y - 5) <= 1

  ##  [1] FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
  ```

# A note on the previous calculation

- Let us produce a small example that will help us explain the computation of the probability. We start by forming a sequence with 10 numbers:

```
Y <- c(6.3, 6.9, 6.6, 3.4, 5.5, 4.3, 6.5, 4.7, 6.1, 5.3)
```

The goal is to compute the proportion of numbers that are in the range $[4, 6]$ (or, equivalently, $\{|Y - 5| \leq 1\}$).

- ```
  abs(Y-5)

  ##  [1] 1.3 1.9 1.6 1.6 0.5 0.7 1.5 0.3 1.1 0.3
  ```

- ```
  abs(Y - 5) <= 1

  ##  [1] FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
  ```

- ```
  mean(abs(Y - 5) <= 1)

  ## [1] 0.4
  ```
  ```

# Probability function

- The probability function of a random variable is defined for any value that the random variable may obtain and produces the **distribution** of the random variable. The probability function may emerge as a relative frequency as in the given example or it may be a result of theoretical modeling.

# Probability function

- The probability function of a random variable is defined for any value that the random variable may obtain and produces the **distribution** of the random variable. The probability function may emerge as a relative frequency as in the given example or it may be a result of theoretical modeling.
- Consider the following probability distribution:

| Value | Probability | Cum.Prob |
|---:|---:|---:|
| 0 | 0.50 | 0.50 |
| 1 | 0.25 | 0.75 |
| 2 | 0.15 | 0.90 |
| 3 | 0.10 | 1.00 |

# Probability function

- The probability function of a random variable is defined for any value that the random variable may obtain and produces the **distribution** of the random variable. The probability function may emerge as a relative frequency as in the given example or it may be a result of theoretical modeling.
- Consider the following probability distribution:

| Value | Probability | Cum.Prob |
|---:|---:|---:|
| 0 | 0.50 | 0.50 |
| 1 | 0.25 | 0.75 |
| 2 | 0.15 | 0.90 |
| 3 | 0.10 | 1.00 |

- What is $\mathrm{P}(Y = 0)$, the probability that $Y$ is equal to 0?:

# Probability function

- The probability function of a random variable is defined for any value that the random variable may obtain and produces the **distribution** of the random variable. The probability function may emerge as a relative frequency as in the given example or it may be a result of theoretical modeling.
- Consider the following probability distribution:

| Value | Probability | Cum.Prob |
|---|---|---|
| 0 | 0.50 | 0.50 |
| 1 | 0.25 | 0.75 |
| 2 | 0.15 | 0.90 |
| 3 | 0.10 | 1.00 |

- What is $\mathrm{P}(Y = 0)$, the probability that $Y$ is equal to 0?:
- What is the probability of $Y$ falling in the interval $[0.5, 2.3]$ ?

# Expectation

- We may characterize the center of the distribution of a random variable and the spread of the distribution in ways similar to those used for the characterization of the distribution of data and the distribution of a population.

# Expectation

- We may characterize the center of the distribution of a random variable and the spread of the distribution in ways similar to those used for the characterization of the distribution of data and the distribution of a population.
- The **expectation** marks the center of the distribution of a random variable. It is equivalent to the data average $\bar{y}$ and the population average $\mu$, which was used in order to mark the location of the distribution of the data and the population, respectively.

# Expectation

- The average of the data can be computed as the weighted average of the values that are present in the data, with weights given by the relative frequency. Specifically, for the data

$$1,\ 1,\ 1,\ 2,\ 2,\ 3,\ 4,\ 4,\ 4,\ 4,\ 4,$$

the mean can be calculated via

$$\begin{aligned}\bar{y} &= \frac{1+1+1+2+2+3+4+4+4+4+4}{11} \\ &= 1 \times \frac{3}{11} + 2 \times \frac{2}{11} + 3 \times \frac{1}{11} + 4 \times \frac{5}{11}\end{aligned}$$

producing the value of $\bar{y} = 2.727$ in both representations.

# Expectation

- The average of the data can be computed as the weighted average of the values that are present in the data, with weights given by the relative frequency. Specifically, for the data

$$1,\ 1,\ 1,\ 2,\ 2,\ 3,\ 4,\ 4,\ 4,\ 4,\ 4,$$

  the mean can be calculated via

$$\begin{aligned}\bar{y} &= \frac{1+1+1+2+2+3+4+4+4+4+4}{11} \\ &= 1 \times \frac{3}{11} + 2 \times \frac{2}{11} + 3 \times \frac{1}{11} + 4 \times \frac{5}{11}\end{aligned}$$

  producing the value of $\bar{y} = 2.727$ in both representations.

- Using a formula, the equality between the two ways of computing the mean is given in terms of the equation:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \sum_{y} \left(y \times (f_y/n)\right) ,$$

  where $f_y$ represents the frequency of $y$ in the data.

# Expectation

- Using a formula, the equality between the two ways of computing the mean is given in terms of the equation:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \sum_{y} \left( y \times (f_y / n) \right) ,$$

where $f_y$ represents the frequency of $y$ in the data.

- The expectation of a random variable is computed in the spirit of the second formulation, and is define via the equation:

$$\mathrm{E}(Y) = \sum_{y} \left( y \times \mathrm{P}(y) \right) .$$

# Variance

- The sample variance ($s^2$) is obtained as the sum of the squared deviations from the average, divided by the sample size ($n$) minus 1:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \; .$$

- A second formulation for the computation of the same quantity is via the use of relative frequencies. The formula for the sample variance takes the form

$$s^2 = \frac{n}{n-1} \sum_{y} \left((y - \bar{y})^2 \times (f_y/n)\right) \; .$$

- In a similar way, the variance of a random variable may be defined via the deviation from the expectation. This deviation is then squared and multiplied by the probability of the value. The multiplications are summed up in order to produce the variance:

$$\mathrm{Var}(Y) = \sum_{y} \left((y - \mathrm{E}(Y))^2 \times \mathrm{P}(y)\right) \; .$$

# Session Info

```
R version 4.0.4 (2021-02-15)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 21.04

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.13.so

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods  
[8] base     

other attached packages:
 [1] cowplot_1.1.0       rio_0.5.16          openintro_2.2.0    
 [4] usdata_0.1.0        cherryblossom_0.1.0 airports_0.1.0     
 [7] oibiostat_0.2.0     DT_0.16             kableExtra_1.2.1   
[10] socviz_1.2          gapminder_0.3.0     here_0.1           
[13] NCStats_0.4.7       FSA_0.8.30          forcats_0.5.1      
[16] stringr_1.4.0       dplyr_1.0.7         purrr_0.3.4        
[19] readr_1.4.0         tidyr_1.1.3         tibble_3.1.3       
[22] ggplot2_3.3.5       tidyverse_1.3.0     knitr_1.33         

loaded via a namespace (and not attached):
 [1] fs_1.5.0           lubridate_1.7.9    webshot_0.5.2      httr_1.4.2        
 [5] rprojroot_2.0.2    backports_1.2.1    utf8_1.2.2         R6_2.5.1          
 [9] DBI_1.1.1          colorspace_2.0-2   withr_2.4.2        tidyselect_1.1.1  
[13] gridExtra_2.3      leaflet_2.0.3      curl_4.3.2         compiler_4.0.4    
[17] cli_3.0.1          rvest_1.0.0        pacman_0.5.1       xml2_1.3.2        
[21] ggdendro_0.1.22    labeling_0.4.2     mosaicCore_0.8.0   scales_1.1.1      
[25] digest_0.6.27      ggformula_0.9.4    foreign_0.8-81     rmarkdown_2.9.7   
[29] pkgconfig_2.0.3    htmltools_0.5.2    highr_0.9          dbplyr_1.4.4      
[33] fastmap_1.1.0      htmlwidgets_1.5.3  rlang_0.4.11       readxl_1.3.1      
[37] rstudioapi_0.13    farver_2.1.0       generics_0.1.0     jsonlite_1.7.2    
[41] crosstalk_1.1.1    zip_2.2.0          car_3.0-9          magrittr_2.0.1    
[45] mosaicData_0.20.1  Matrix_1.3-2       Rcpp_1.0.7         munsell_0.5.0     
[49] fansi_0.5.0        abind_1.4-5        lifecycle_1.0.0    stringi_1.7.3     
[53] carData_3.0-4      MASS_7.3-53.1      plyr_1.8.6         ggstance_0.3.4    
[57] grid_4.0.4         blob_1.2.1         ggrepel_0.8.2      crayon_1.4.1      
[61] lattice_0.20-41    haven_2.3.1        splines_4.0.4      hms_1.0.0         
[65] pillar_1.6.2       reprex_0.3.0       glue_1.4.2         evaluate_0.14     
[69] data.table_1.14.0  modelr_0.1.8       vctrs_0.3.8        tweenr_1.0.1      
```