Lesson 1:
Chapter 1

Caleb Moxley

BSC Mathematics

31 August 15

§1.1 Data - The Basics

Definition (case)

A **case** is an object within a data set.

Example (case)

If our data set describes the heights of students, then a case within our data set is a student.

§1.1 Data - The Basics

Definition (case)

A **case** is an object within a data set.

Example (case)

If our data set describes the heights of students, then a case within our data set is a student.

Definition (label)

A **label** is a variable used in a data set to distinguish between different cases.

§1.1 Data - The Basics

Definition (case)

A **case** is an object within a data set.

Example (case)

If our data set describes the heights of students, then a case within our data set is a student.

Definition (label)

A **label** is a variable used in a data set to distinguish between different cases.

Example (Cases)

If our data set describes the heights of students, then a label for a particular case may be a student's name or perhaps her student ID number.

§1.1 Data - The Basics

Definition (variable)

A **variable** is a characteristic of a case.

§1.1 Data - The Basics

Definition (variable)

A **variable** is a characteristic of a case.

Example (variable)

If our data set describes the heights of students, then the only non-label variable is the height.

§1.1 Data - The Basics

Definition (variable)

A **variable** is a characteristic of a case.

Example (variable)

If our data set describes the heights of students, then the only non-label variable is the height.

Is a label a variable?

§1.1 Data - The Basics

Definition (distribution)

A **distribution** of a variable gives all possible values of a variable and how often these values occur.

§1.1 Data - The Basics

Definition (distribution)

A **distribution** of a variable gives all possible values of a variable and how often these values occur.

Example (distribution)

The following is a (frequency) distribution of heights in a classroom:

Height (cm)	Frequency
150-159	3
160-169	6
170-179	9
180-189	6
190-199	1

§1.1 Data - The Basics

There are two basic variable types: categorical (qualitative) and numerical (quantitative) variables.

§1.1 Data - The Basics

There are two basic variable types: categorical (qualitative) and numerical (quantitative) variables.

A categorical variable takes on one of several (often non-measurement) values coming from a list of possible categories.

§1.1 Data - The Basics

There are two basic variable types: categorical (qualitative) and numerical (quantitative) variables.

A categorical variable takes on one of several (often non-measurement) values coming from a list of possible categories.

A quantitative variable takes numerical values which **must measure something**. Arithmetic makes sense only on quantitative variables.

§1.1 Data - The Basics

There are two basic variable types: categorical (qualitative) and numerical (quantitative) variables.

A categorical variable takes on one of several (often non-measurement) values coming from a list of possible categories.

A quantitative variable takes numerical values which **must measure something**. Arithmetic makes sense only on quantitative variables.

Can numbers be categorical variables?

§1.1 Data - The Basics

A caveat: Make sure that the variables chosen measure what you're actually interested in studying.

§1.1 Data - The Basics

A caveat: Make sure that the variables chosen measure what you're actually interested in studying.

What pitfall might arise here? A researcher is interested in crime in cities, so she obtains a data set which gives the number of crimes in each of city. She reports on the most dangerous cities based on the number of crimes in each city.

§1.1 Data - The Basics

A caveat: Make sure that the variables chosen measure what you're actually interested in studying.

What pitfall might arise here? A researcher is interested in crime in cities, so she obtains a data set which gives the number of crimes in each of city. She reports on the most dangerous cities based on the number of crimes in each city.

She should be focusing on **crime rates** rather than simply the number of crimes!

§1.2 Displaying Distributions Graphically

Gaining a basic understanding of a data set involves exploratory data analysis - the major component of which is graphical summary and numerical summary.

§1.2 Displaying Distributions Graphically

Gaining a basic understanding of a data set involves exploratory data analysis - the major component of which is graphical summary and numerical summary.

We begin with a graphical summaries for categorical and quantitative data.

§1.2 Displaying Distributions - Categorical Data

A basic image of the distribution of a categorical data set is easily given by a bar graph or a pie chart.

§1.2 Displaying Distributions - Categorical Data

A basic image of the distribution of a categorical data set is easily given by a bar graph or a pie chart. We can use our previous example.

§1.2 Displaying Distributions - Categorical Data

A basic image of the distribution of a categorical data set is easily given by a bar graph or a pie chart. We can use our previous example.

Height (cm)	Frequency
150-159	3
160-169	6
170-179	9
180-189	6
190-199	1

What would the distribution look like if we converted it to percentages?

§1.2 Displaying Distributions - Categorical Data

A basic image of the distribution of a categorical data set is easily given by a bar graph or a pie chart. We can use our previous example.

Height (cm)	Frequency
150-159	$\frac{3}{25} = 12\%$
160-169	6
170-179	9
180-189	6
190-199	1

What would the distribution look like if we converted it to percentages?

§1.2 Displaying Distributions - Categorical Data

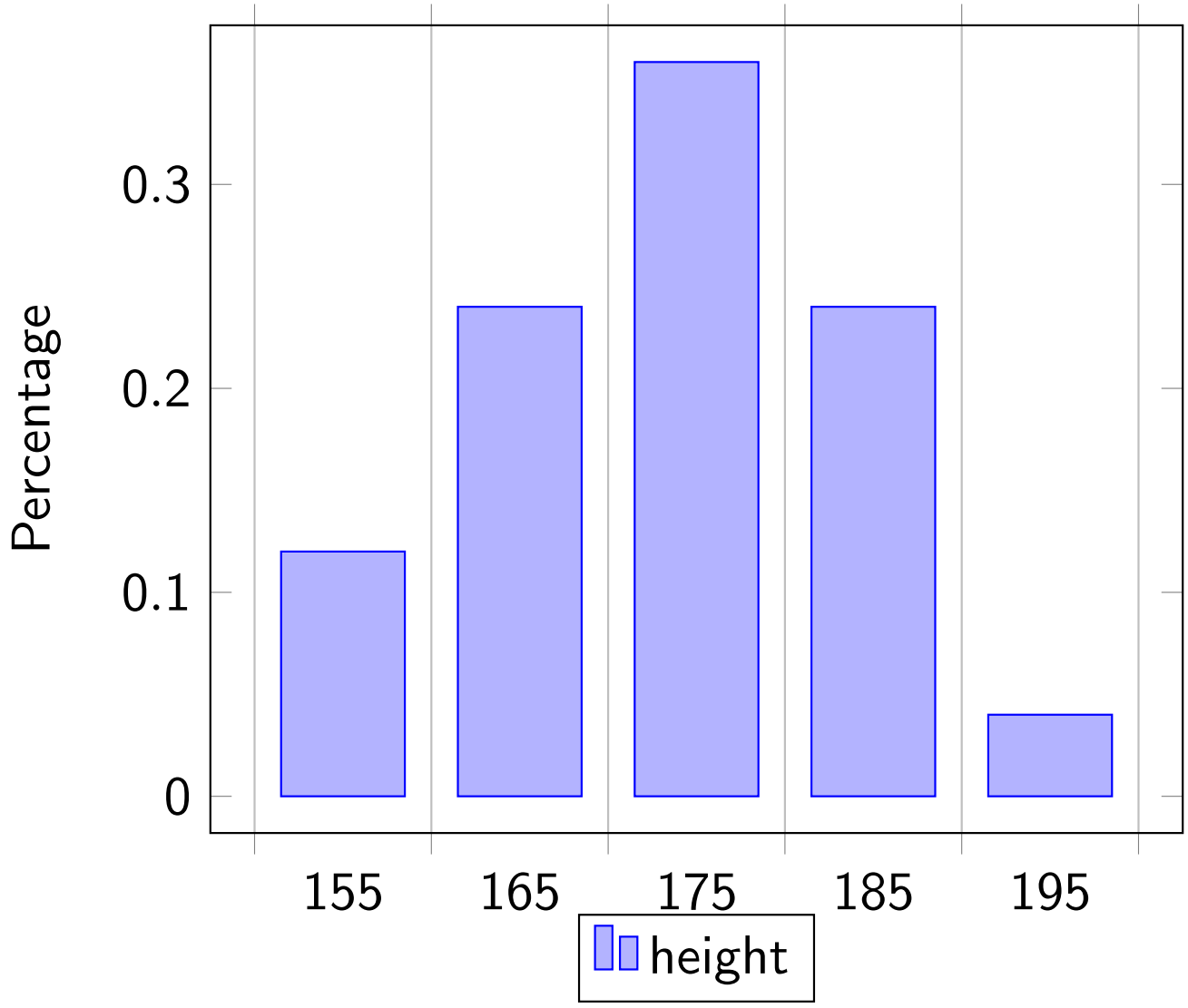
A basic image of the distribution of a categorical data set is easily given by a bar graph or a pie chart. We can use our previous example.

Height (cm)	Frequency
150-159	12%
160-169	24%
170-179	36%
180-189	24%
190-199	4%

What would the distribution look like if we converted it to percentages?

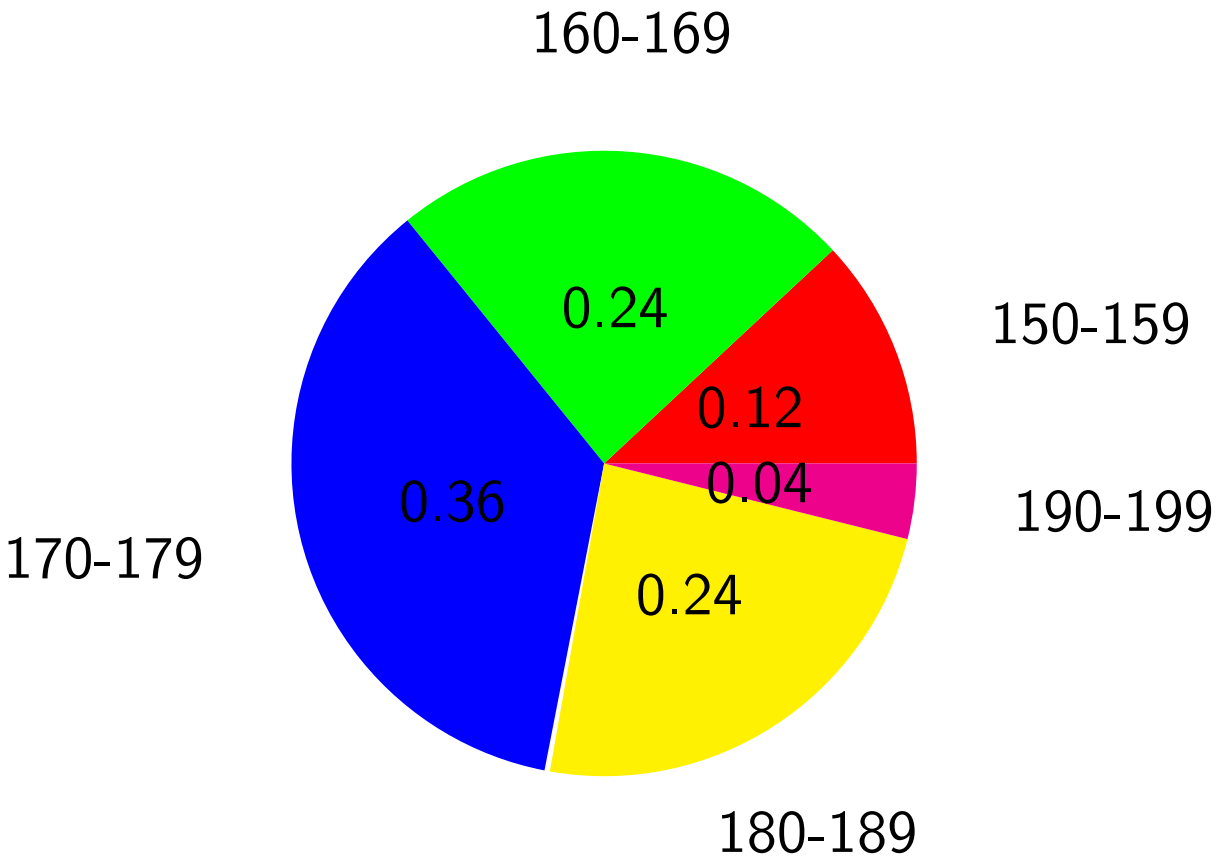
§1.2 Displaying Distributions - Categorical Data

§1.2 Displaying Distributions - Categorical Data



§1.2 Displaying Distributions - Categorical Data

§1.2 Displaying Distributions - Categorical Data



§1.2 Displaying Distributions - Quantitative Data

Let's say the actual data that we had of heights of students in a class was the following: 151, 155, 156, 160, 161, 165, 165, 167, 169, 170, 170, 171, 175, 177, 177, 177, 177, 179, 185, 186, 187, 188, 188, 189, 199. We can summarize this (truly) quantitative data using a **stemplot**:

§1.2 Displaying Distributions - Quantitative Data

Let's say the actual data that we had of heights of students in a class was the following: 151, 155, 156, 160, 161, 165, 165, 167, 169, 170, 170, 171, 175, 177, 177, 177, 177, 179, 185, 186, 187, 188, 188, 189, 199. We can summarize this (truly) quantitative data using a **stemplot**:

15	156
16	
17	
18	
19	

§1.2 Displaying Distributions - Quantitative Data

Let's say the actual data that we had of heights of students in a class was the following: 151, 155, 156, 160, 161, 165, 165, 167, 169, 170, 170, 171, 175, 177, 177, 177, 177, 179, 185, 186, 187, 188, 188, 189, 199. We can summarize this (truly) quantitative data using a **stemplot**:

15	156
16	015579
17	001577779
18	567889
19	9

§1.2 Displaying Distributions - Quantitative Data

We may use **back-to-back stemplots** to compare two related distributions. Say a * indicates a male student's height and no * indicates a female: 151*, 155, 156, 160, 161*, 165, 165, 167, 169, 170*, 170*, 171*, 175, 177, 177, 177*, 177*, 179*, 185, 186*, 187, 188, 188*, 189*, 199*:

Male		Female
1	15	56
1	16	05579
977100	17	577
986	18	578
9	19	

§1.2 Displaying Distributions - Quantitative Data

You can also refine a stemplot by only allowing digits to go between 0-4 and 5-9 in each stem. This is called **splitting**. You can also **trim** a stemplot by dropping the final digit of a piece of data.

§1.2 Displaying Distributions - Quantitative Data

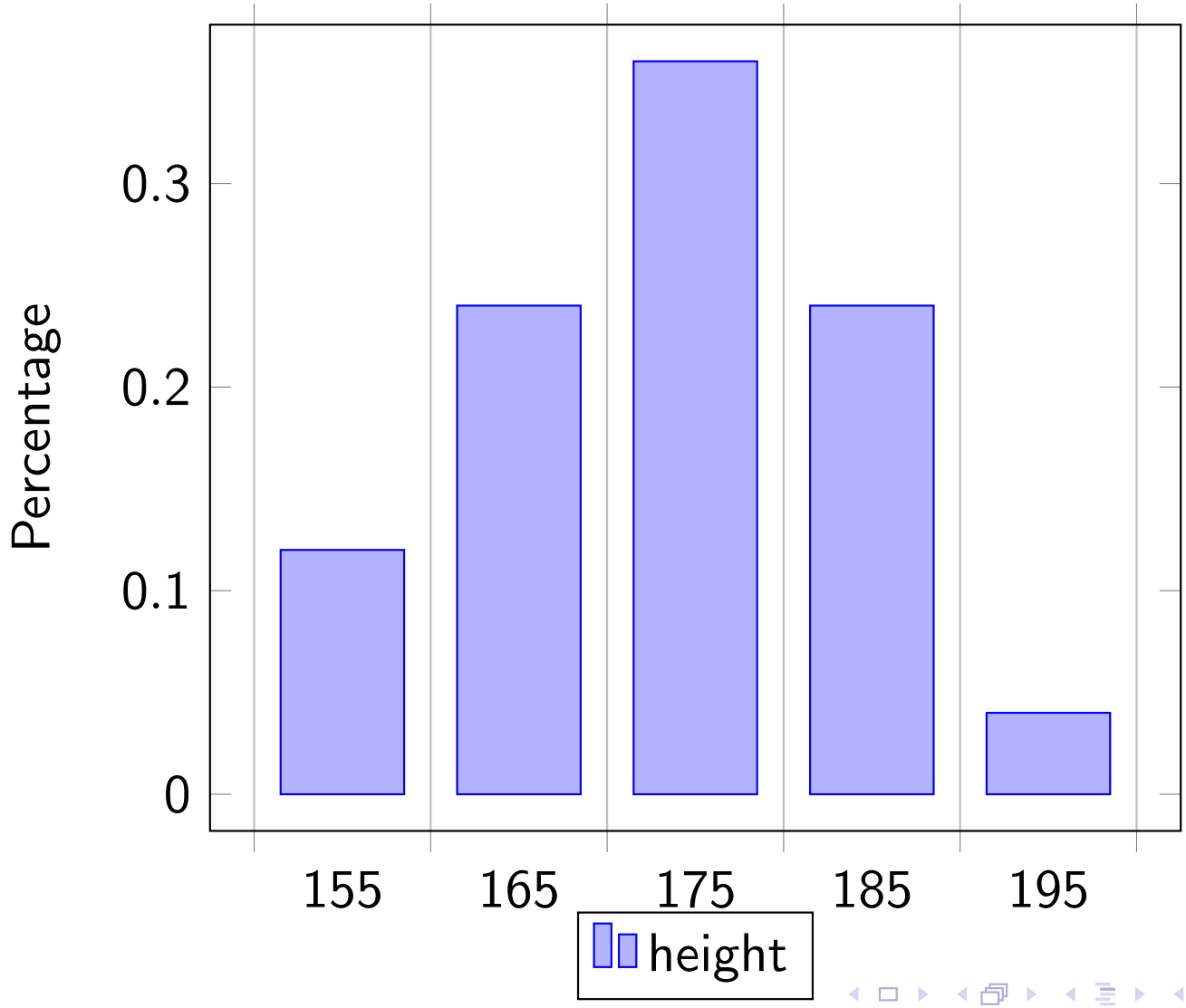
You can also refine a stemplot by only allowing digits to go between 0-4 and 5-9 in each stem. This is called **splitting**. You can also **trim** a stemplot by dropping the final digit of a piece of data. Which gives a more detailed stemplot? Which gives a less detailed stemplot?

§1.2 Displaying Distributions - Quantitative Data

You can also use a histogram to graphically represent data.

§1.2 Displaying Distributions - Quantitative Data

You can also use a histogram to graphically represent data.

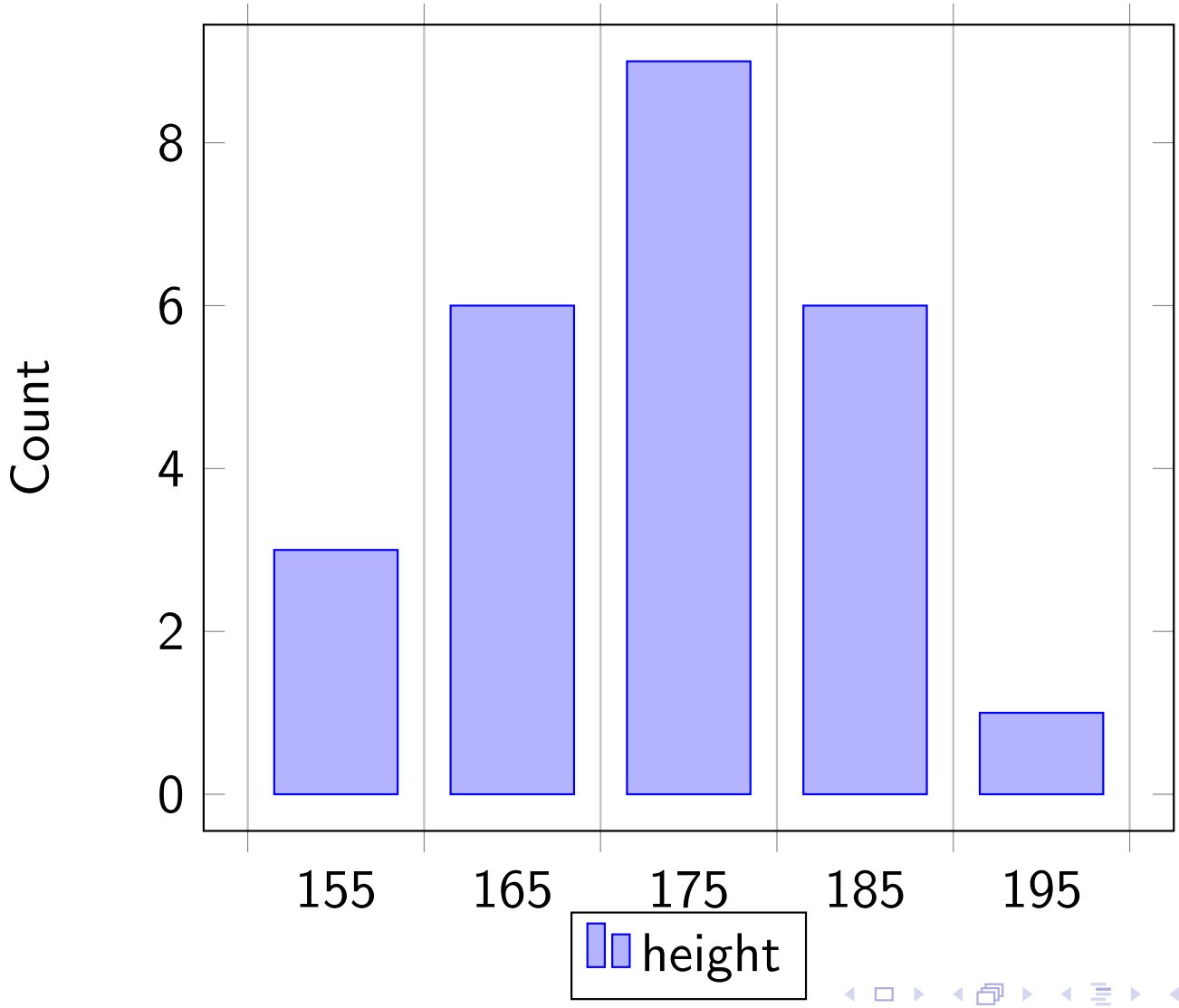


§1.2 Displaying Distributions - Quantitative Data

You can use counts rather than percentages as well.

§1.2 Displaying Distributions - Quantitative Data

You can use counts rather than percentages as well.



§1.2 Displaying Distributions - Quantitative Data

Note: You should actually draw histograms **without any spaces** between the bars. I have not done that in these notes. Some programs simply draw histograms as if they are bar graphs.

§1.2 Displaying Distributions - Quantitative Data

Note: You should actually draw histograms **without any spaces** between the bars. I have not done that in these notes. Some programs simply draw histograms as if they are bar graphs. What types of variables must be graphically represented as bar graphs rather than histograms?

§1.2 Displaying Distributions - Quantitative Data

Once a variable has been graphically represented, you can see features of the distribution.

§1.2 Displaying Distributions - Quantitative Data

Once a variable has been graphically represented, you can see features of the distribution.

1. skewness (left- or right-skewness)
2. mode(s)
3. shape
4. tails
5. deviations/outliers
6. symmetry

You should be able to identify these basic features in a graph.

§1.2 Displaying Distributions - Quantitative Data

Time plots plot observations against the time at which they were observed. Time is always the horizontal axis variable. We won't discuss time plots extensively.

§1.3 Numerical Descriptions of Data

Comparing distributions can be difficult to do graphically, so numerical descriptions of distributions are essential. We're usually interested in **spread/variation** within a distribution and its **center**.

§1.3 Numerical Descriptions of Data

Comparing distributions can be difficult to do graphically, so numerical descriptions of distributions are essential. We're usually interested in **spread/variation** within a distribution and its **center**. Mean and median are two typical measures of center while quartiles and standard deviation/variance are the typical measures of spread.

§1.3 Numerical Descriptions of Data

Definition (mean)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \sum x_i$$

Definition (median)

If the data are ordered from smallest to lowest, then

$$M = x_{\frac{n+1}{2}}$$

if n is odd and

$$M = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

if n is even.

§1.3 Numerical Descriptions of Data

Definition (mean)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \sum x_i$$

Definition (median)

If the data are ordered from smallest to lowest, then

$$M = x_{\frac{n+1}{2}}$$

if n is odd and

$$M = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

if n is even.

Note: Medians are robust (resistant) against outliers. Means are

§1.3 Numerical Descriptions of Data

Example (mean)

$$\frac{151 + 155 + 156 + 160 + \cdots + 188 + 188 + 189 + 199}{25} =$$

§1.3 Numerical Descriptions of Data

Example (mean)

$$\frac{151 + 155 + 156 + 160 + \cdots + 188 + 188 + 189 + 199}{25} = \frac{4344}{25}$$

$= 173.76$

Example (median)

Since $n = 25$, we take the 13th largest number from the list for the median: 151, 155, 156, 160, 161, 165, 165, 167, 169, 170, 170, 171, 175, 177, 177, 177, 177, 179, 185, 186, 187, 188, 188, 189, 199.

§1.3 Numerical Descriptions of Data

Example (mean)

$$\frac{151 + 155 + 156 + 160 + \cdots + 188 + 188 + 189 + 199}{25} = \frac{4344}{25}$$

$= 173.76$

Example (median)

Since $n = 25$, we take the 13th largest number from the list for the median: 151, 155, 156, 160, 161, 165, 165, 167, 169, 170, 170, 171, 175, 177, 177, 177, 177, 179, 185, 186, 187, 188, 188, 189, 199. Thus, $M = 175$.

§1.3 Numerical Descriptions of Data

Definition (quartiles)

The first **quartile** (Q_1) of a data set is the median of the values to the left of the median. $Q_2 = M$ and Q_3 is the median of the values to the right of the median.

Definition (interquartile range)

$$IQR = Q_3 - Q_1$$

We often use the IQR to label suspected outliers if they lie $1.5 \cdot IQR$ above the third quartile or $1.5 \cdot IQR$ below the first quartile.

§1.3 Numerical Descriptions of Data

Find the five-number summary (Min, Q_1, M, Q_3, Max) of the heights of the class and draw a boxplot!

151, 155, 156, 160, 161, 165, 165, 167, 169, 170, 170, 171, 175, 177, 177, 177, 177, 179, 185, 186, 187, 188, 188, 189, 199.

§1.3 Numerical Descriptions of Data

Find the five-number summary (Min, Q_1, M, Q_3, Max) of the heights of the class and draw a boxplot!

151, 155, 156, 160, 161, 165, 165, 167, 169, 170, 170, 171, 175, 177, 177, 177, 177, 179, 185, 186, 187, 188, 188, 189, 199.

1. The minimum is 151.
2. Q_1 is the median of the first twelve numbers: 165.
3. $M = 175$.
4. Q_3 is the median of the last twelve numbers: 185.5.
5. The maximum is 199.

§1.3 Numerical Descriptions of Data

Find the five-number summary (Min, Q_1, M, Q_3, Max) of the heights of the class and draw a boxplot!

151, 155, 156, 160, 161, 165, 165, 167, 169, 170, 170, 171, 175, 177, 177, 177, 177, 179, 185, 186, 187, 188, 188, 189, 199.

1. The minimum is 151.
2. Q_1 is the median of the first twelve numbers: 165.
3. $M = 175$.
4. Q_3 is the median of the last twelve numbers: 185.5.
5. The maximum is 199.

Thus, $IQR = 20.5$ and there are no suspected outliers.

§1.3 Numerical Descriptions of Data

Definition (variance)

The **variance** (s^2) of a data set is

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum(x_i^2) - n\bar{x}^2}{n-1}.$$

Definition (standard deviation)

The **standard deviation** (s) of a data set is

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum(x_i^2) - n\bar{x}^2}{n-1}}.$$

Standard deviations and variances are not resistant against outliers.

§1.3 Numerical Descriptions of Data

Compute the variance and standard deviation of the heights of students in a class.

§1.3 Numerical Descriptions of Data

Compute the variance and standard deviation of the heights of students in a class.

- $\bar{x} = 173.76$.
- $\sum(x_i)^2 = 758406$.
- $n - 1 = 24$.

§1.3 Numerical Descriptions of Data

Compute the variance and standard deviation of the heights of students in a class.

- $\bar{x} = 173.76$.
- $\sum(x_i)^2 = 758406$.
- $n - 1 = 24$.

So we have $s^2 = \frac{758406-(25)(173.76)^2}{24} = 149.69$ and $s = 12.2348$.

§1.3 Numerical Descriptions of Data

If you want to change a random variable via a linear transformation, then the mean and variance changes, too! Say we want to change heights in centimeters to heights with a 10cm hat on given in meters, then 151 would be $\frac{151+10}{100}$, i.e. we transform via the linear transformation $x_{\text{new}} = 0.1 + 0.01x$.

§1.3 Numerical Descriptions of Data

If you want to change a random variable via a linear transformation, then the mean and variance changes, too! Say we want to change heights in centimeters to heights with a 10cm hat on given in meters, then 151 would be $\frac{151+10}{100}$, i.e. we transform via the linear transformation $x_{\text{new}} = 0.1 + 0.01x$.

The mean changes via the same linear transformation.

§1.3 Numerical Descriptions of Data

If you want to change a random variable via a linear transformation, then the mean and variance changes, too! Say we want to change heights in centimeters to heights with a 10cm hat on given in meters, then 151 would be $\frac{151+10}{100}$, i.e. we transform via the linear transformation $x_{\text{new}} = 0.1 + 0.01x$.

The mean changes via the same linear transformation. We have

$$\bar{x}_{\text{new}} = 0.1 + 0.01(173.76) = 1.8376.$$

However, the variance only changes based on the **slope** of the linear transformation

$$s^2_{\text{new}} = 0.01^2(149.69) = 0.014969.$$

§1.4 Density Curves and Normal Distributions

You can fit a smooth curve to the irregular bars of a histogram to create a **density curve**.

Definition (density curve)

A **density curve** is a curve that

1. is always above or on the horizontal axis and
2. has an area of exactly one underneath it

§1.4 Density Curves and Normal Distributions

You can fit a smooth curve to the irregular bars of a histogram to create a **density curve**.

Definition (density curve)

A **density curve** is a curve that

1. is always above or on the horizontal axis and
2. has an area of exactly one underneath it

These curves describe the overall pattern of a distribution. The area under the curve between two values is the proportion of values in the distribution which fall between those bounds.

§1.4 Density Curves and Normal Distributions

A **normal density curve** is a special density curve characterized by a bell shape, no skewness, and symmetry.

§1.4 Density Curves and Normal Distributions

A **normal density curve** is a special density curve characterized by a bell shape, no skewness, and symmetry.

Normal distribution frequencies

1. start small,
2. increase to a few high frequency counts,
3. taper back to small counts again,
4. and are symmetric.

§1.4 Density Curves and Normal Distributions

When we talk about the standard deviation and mean of density curves, we use the symbols σ and μ respectively. The normal distribution has the following properties:

1. 68% of observations fall within σ of μ,
2. 95% of observations fall within 2σ of μ, and
3. 99.7% of observations fall within 3σ of μ.

§1.4 Density Curves and Normal Distributions

Definition (standard score)

If x is an observation from a distribution with mean μ and standard deviation σ, then its **standard score** z is given by

$$z = \frac{x - \mu}{\sigma}.$$

Note: You can convert any normal distribution with mean μ and standard deviation σ into a **standard normal distribution** by converting all values to their standard score! The new mean and standard deviation would be 0 and 1, respectively.

§1.4 Density Curves and Normal Distributions

Look at the stemplot of the heights of students in a class. Does it appear to come from a normally distributed population?

15	156
16	015579
17	001577779
18	567889
19	9

§1.4 Density Curves and Normal Distributions

Look at the stemplot of the heights of students in a class. Does it appear to come from a normally distributed population?

15	156
16	015579
17	001577779
18	567889
19	9

What's the standard score of 151?

§1.4 Density Curves and Normal Distributions

Look at the stemplot of the heights of students in a class. Does it appear to come from a normally distributed population?

15	156
16	015579
17	001577779
18	567889
19	9

What's the standard score of 151? It's $\frac{151-173.76}{12.2348} = -1.86$.

§1.4 Density Curves and Normal Distributions

Please review how to use a standard normal table to calculate normal probabilities on your own. Most of you have probably already done this! It should be easy review. See pages 64-68.