|  |
| --- |
| Versions: Andrew McConnell Dec.04/08  Bonnie Johnston Dec.12/08//Jan21/08  Andrew McConnell May 20/09 |

|  |
| --- |
| **Module Names** are in red  **(h1)Topic Names** are in pink  **Headings and content** for the pages are in black  **h1** is heading 1**, h2** is heading 2, **h3** is heading 3, **h4** is heading 4  **Sub-modules** (sections within modules) are in blue  **Quiz title** in green (as quicklink in TOC)  **QL Quicklinked Page links within course** to page Topic Name are in purple  **QL Quicklinked External links** also in purple  **Assignments for dropbox** are in orange  SME/author attention needed highlighted in yellow |

**Module 2 Summary Statistics**

**Getting Started with Module 2**

**(h1)Introduction to Summary Statistics**

**(h2) Large Medical Research Studies**

The Women’s Health Initiative (WHI) study is, to quote their website,

a long-term national health study that focuses on strategies for preventing heart disease, breast and colorectal cancer and fracture in postmenopausal women.

See www.whi.org

This study involved over 160,000 women aged 50-79 and followed this cohort from 1993 to 2005 with a new extension of 5 years continuing to 2010. The amount of data gathered is quite phenomenal since even just the weight of each woman constitutes 160,000 data points. One common method of showing people data and explaining its meaning (called summarizing data) is that of just showing the raw data but as you may imagine this is a silly concept when dealing with such a large set of data such as is being gathered in the WHI study. Statistics come to the rescue with descriptive statistics methods. For example, an analysis of bone mineral density when the subjects were either on calcium with a vitamin D supplement or not, had two study groups with about 18,000 subjects in each. The following plot was used to summarize the differences between the two groups based on their bone mineral density.

Taken from *New England Journal of Medicine* **354**, p 669

Notice that despite having literally thousands of bone mineral density measurements the results can be conveniently summarized on a single plot whose meaning can be easily interpreted. Note that the word “mean” is a statistics term for the “average”, where we add all the measurements together and divide by the total number of measurements – more about this later.

The purpose of this module is to get you comfortable with using and understanding statistics as a way to present results and summarize larger sets of data.

**(h2) Learning Objectives**

By the end of this module, you will be able to:

* define and set up a class, class boundaries, class limits, and class marks.
* classify data by type.
* differentiate between frequency, cumulative frequency, relative frequency, and cumulative relative frequency plots.
* choose an appropriate graphical method to display data for summary and exploratory purposes.
* calculate the following measures of central tendencies: mean, median, and mode.
* calculate the measures of positions: quartiles, percentiles, and rank.
* calculate measures of dispersion for a data set: variance, standard deviation, and range.
* differentiate between the calculation of these measures for a population and a sample.
* understand the significance of standard deviation as applied to the Empirical Rule.

**Pre-module 2 Activity**

(multiple choice)

In a recent study of … at Vancouver General Hospital.. Do you think the population of interest in this study is

a) the group of people who … and were involved in the study at Vancouver General Hospital

b) All people who …. at Vancouver General Hospital

c) All people who…

**(h1) How to Proceed through Module 2**

There are three sections in this module. Work your way through each section’s online and textbook readings, examples, and practice problems, and try the self tests.

**(h2) Tables and Graphical Methods**

* Read Chapter 2 Sections 1, 2 and 3 in your textbook.
* Read **Graphical Summary of Data**
* Do Chapter 2 Section 2 Problems # 1, 5, 15
* Do Chapter 2 Section 3 Problems # 1, 5, 8, 9
* Do Module 2 Self Test 1

**(h2) Numerical Summary Statistics**

* Read Chapter 2 Sections 4, 5 and 6 in your textbook.
* Read **Numerical Summary of Data**

In this section you should practice using your calculator to get the mean and standard deviation; it will be expected that you have this available to you in a test. Don’t use a computer program as it will be unavailable during a test but you can use a computer to check your results.

* Do Chapter 2 Section 4 Problems # 1,3 7, 13, 21
* Do Chapter 2 Section 5 Problems # 3, 7, 9, 13, 21, 23
* Do Chapter 2 Section 6 Problems # 1, 3, 7, 11, 23
* Do Module 2 Self Test 2

**(h2) Exploratory Data Analysis**

* Read Chapter 2 Sections 7 in your textbook.
* Read **Exploring the Data**
* Do Chapter 2 Section 7 Problems # 3, 5, 9
* Do Module 2 Self Test 3

**(h2) Assignments**

Once you’ve finished the module work, complete Assignment 2 and submit it to your instructor via the Dropbox online.

**(h2) Advanced Readings**

Here are some advanced readings about the topics in this module. The following are optional

* Variance versus Standard Deviation
* Why is it ?

**Graphical Summary of Data**

**(h1)Graphical Summary of Data**

**Calculation** **Note**: Feel free to use Statdisk, Excel™ or another program to make your plots and graphs but be very careful if you are using a general program such as Excel™ , because some default plots are not very good statistical plots. One thing that is important is that you make sure you fully understand how such plots are created so that you can do them yourself if needed in this course. Remember that you will not have a computer in the test with you.

**.**

Of the many types of plots seen in this section, probably the most common in research are the histogram, the x-y (scatter) plot and the “box plot”.

A distinction can be made between different types of bar graphs and indeed some feel that a histogram should not be called a bar graph. While bar graph can use discrete categories (such as male or female) a histogram has an *x*-axis that is a continuous range of values that are then divided up into classes.

|  |  |
| --- | --- |
| a) Histogram | b) Bar Chart of Discrete Categories |

One issue with histograms is the choosing of the size of the class width. To see why this is an issue try either of the following applets and play with the class width sizes.

* **<a href=”http://www.stat.sc.edu/~west/javahtml/Histogram.html”> Histogram Class Width Applet 1</a>**
* **<a href=”http://www.shodor.org/interactivate/activities/histogram/”> Histogram Class Width Applet 2</a>**

As you can see, the choice of the width can make a critical difference how one describes data that is collected. e.g. Is it mounded in the center, does it have two different peaks, etc.

We will return to the *x*-*y* scatter plots near the end of the course in the topic of linear regression so we will not spend much time on them right now. Also, since the box plot is of most use in comparing different sets of similar data, it will be addressed in the Exploratory Data Analysis section later in the module.

**(h2) Chart abuse**

The textbook mentions the inappropriateness of using a pie chart. One reason for this dislike is that the pie chart requires the reader to make estimates of relative areas which can be difficult. With the advent of spreadsheet programs a new problem has arisen with this, three dimensional (3D) pie charts. By making what is fundamentally a 3D chart into a pseudo 3D chart the difficulty of estimating relative areas has increased. For example the following chart was taken from a paper in Clinical Cardiology,volume**31**. While the areas which are very different are pretty clear, the three sections with 1% given a visual impression of being different due to the oblique angle of viewing.

This difficulty is by no means limited to pie charts but also applies to the unnecessary use of “3D” in bar graphs where is suddenly can become difficult to determine the correct height of a bar due to the 3D effect.

**(h2)Survival Plots**

Another plot occasionally seen in medical research papers is called a “survival plot”. Although the creation of these curves is beyond the scope of the present course, their interpretation is fairly simple. The curve gives an estimate of the probability of survival (*y* axis) – either mortality or disease-free – to the time indicated on the *x* axis. The plot shown below is a Kaplan-Meier survival curve and allows calculation of probability of survival even if subjects drop out of the study part way. The example shown below is taken from “*Effect of Statins and White Blood Cell Count on Mortality in Patients with Ischemic Left Ventricular Dysfunction Undergoing Percutaneous* *Coronary Intervention”*, M. Lipinski *et al.*, Clinical Cardiology, **29**,p36. Looking at the plot one can see that the probability of surviving to 5.5 years taking statins is approximately 0.90 (or 90%).

**(h1)Example of Graphical Summary**

The numbers are the time to return to resting heart rate in minutes for 30 kinesiology volunteers after a very mild exercise.

**Example Data Set 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Time to Return to Resting Heart Rate (min)** | | | | | |
| 1.9 | 2.4 | 2.6 | 2.9 | 2.9 | 3.2 |
| 2.0 | 2.5 | 2.7 | 2.9 | 3.0 | 3.2 |
| 2.2 | 2.6 | 2.7 | 2.9 | 3.0 | 3.4 |
| 2.2 | 2.6 | 2.7 | 2.9 | 3.1 | 3.4 |
| 2.4 | 2.6 | 2.8 | 2.9 | 3.2 | 3.4 |

**(h2) Discussion**

The first step in condensing a large data set is to subdivide it into classes. Boundaries are simply the dividing lines between successive classes. The boundaries of the classes are typically chosen to be offset from the data by half of the smallest data increment used so that no ambiguity or overlap occurs (e.g., for the data above, the smallest data increment is 0.1; therefore, boundaries are chosen ending in 0.1/2 = 0.05).

The class boundaries for the example data could be constructed as follows although this is by no means the only way to break the data up into classes.

|  |  |
| --- | --- |
| **Class** | **Class Boundaries** |
| 1  2  3  4  5  6  7  8 | 1.85 to 2.05  2.05 to 2.25  2.25 to 2.45  2.45 to 2.65  2.65 to 2.85  2.85 to 3.05  3.05 to 3.25  3.25 to 3.45 |

Note that these could have been as easily listed as 1.9 to < 2.1, 2.1 to < 2.3 ... 3.3 to < 3.5. Either method of specifying the boundary is correct.

It is more usual when constructing a frequency table, particularly when it is to be used in published material, to have a column marked class limits and omit the column marked class boundaries. Class limits have the same decimal placement as the underlying raw data, but like boundaries, allow no overlap or ambiguity to exist. The smallest and largest values that can fall into a given class are referred to as its class limits. For the example data set 1, the correct class limits are as follows:

|  |  |
| --- | --- |
| **Class** | **Class Limits** |
| 1  2  3  4  5  6  7  8 | 1.9–2.0  2.1–2.2  2.3–2.4  2.5–2.6  2.7–2.8  2.9–3.0  3.1–3.2  3.3–3.4 |

For further practice and assistance in understanding the relationship between class limits and class boundaries, refer to the sample problems and complete solutions at the end of this lesson.

We can then construct the whole frequency table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class** | **Class Boundaries** | **Class**  **Limits** | **Class**  **Frequency** | **Cumulative**  **Frequency** | **Class**  **Mark** |
| 1  2  3  4  5  6  7  8 | 1.85 to 2.05  2.05 to 2.25  2.25 to 2.45  2.45 to 2.65  2.65 to 2.85  2.85 to 3.05  3.05 to 3.25  3.25 to 3.45 | 1.9–2.0  2.1–2.2  2.3–2.4  2.5–2.6  2.7–2.8  2.9–3.0  3.1–3.2  3.3–3.4 | 2  2  2  5  4  8  4  3 | 2    4    6  11  15  23  27  30 | 1.95  2.15  2.35  2.55  2.75  2.95  3.15  3.35 |

From this we can construct a number of different plots, a frequency polygon, an ogive

|  |  |
| --- | --- |
| K:\GRAPHICS\HSC\HSC00000\HSC00683.TIF | K:\GRAPHICS\HSC\HSC00000\HSC00684.TIF |

and a histogram (both frequency and relative frequency histograms are shown here).

|  |  |
| --- | --- |
|  |  |

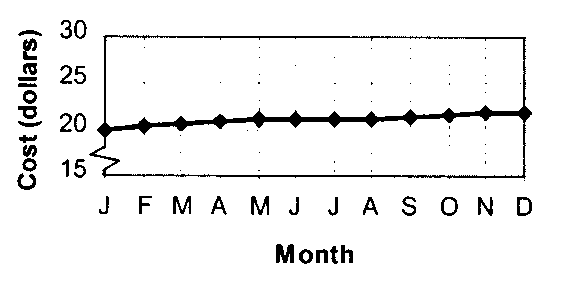
**(h2)A word of caution.**

Pictures provide an excellent way to distort the truth. For example, recall the pictures of a man or woman taken before and after a certain dieting method is used. You will notice in the before picture the person is often shown wearing drab clothes and solemn expression; while in the after picture the person is certainly thinner, but this effect is exaggerated by smartly styled clothes, a new hairdo, and a radiant smile. We generally take such mini-frauds for granted. However, statistical pictures, such as histograms, bar graphs, and line charts, can also be used to distort the truth, but in more subtle ways. People, in general, tend to shy away from large masses of numbers; most people are grateful to simply glance at a colorful or amusing concise drawing of the underlying facts. It is in this glance that the reader carries away a visual image of what they think underlies the picture. In order to see how this deception can work consider the following example.

|  |  |
| --- | --- |
| K:\GRAPHICS\HSC\HSC00000\HSC00685.TIF  (a) | K:\GRAPHICS\HSC\HSC00000\HSC00686.TIF  (b) |
| K:\GRAPHICS\HSC\HSC00000\HSC00687.TIF  (c) | K:\GRAPHICS\HSC\HSC00000\HSC00688.TIF  (d) |

All of the above graphs depict the same data set indicating the rise in cost of a certain medication over the period of one year. Graphs (a) and (c) have the same effective scaling; however, the scale starts at zero for graph (c) and at $15 for graph (a). By having the scale start at $15 in graph (a), it tends to “hide” the base cost of $20. Graphs (b) and (d) both tend to emphasize the absolute change in price of $2, once again perhaps losing sight that this is a change from $20 to $22. Such techniques are particularly deceiving when placed in television adds where you are not given time to fully analyze what is being shown.

If it is necessary to start your horizontal or vertical axes at a place other than zero, be sure and at least draw the reader’s attention to it by making a clear jog on the axes as shown in the following graph.



Now try the following practice problems and compare your results with the solutions provided.

**(h1)Practice Problem 1**

A study of the body’s ability to absorb lead is conducted by using radioactive tracer techniques. The percentages absorbed, to the nearest 0.1%, are grouped into a table having the classes 12.5–13.4, 13.5–14.4, 14.5–15.4, 15.5–16.4, 16.5–17.4, and 17.5–18.4.

Determine the:

A. lower class limits.

B. upper class limits.

C. length of the class interval or class width.

D. lower class boundaries.

E. upper class boundaries.

**(h1)Solution to Practice Problem 1**

1. 12.5, 13.5, 14.5, 15.5, 16.5, 17.5
2. 13.4, 14.4, 15.4, 16.4, 17.4, 18.4
3. To find the class width, simply subtract successive lower boundaries, or successive upper boundaries, or successive lower limits, or successive upper limits, or the difference between class boundaries of the same class. Here the class width is 1.0.

***Note***: Subtracting the lower class limit from the upper class limit of the same class does ***NOT*** give you the class width.

1. The upper boundaries are: 13.45, 14.45, 15.45, 16.45, 17.45, and 18.45.
2. The lower boundaries are: 12.45, 13.45, 14.45, 15.45, 16.45, and 17.45.

**(h1)Practice Problem 2**

Measurements of the boiling point of a substance, measured to the nearest degree Celsius, vary from 136° to 168°. Indicate the limits of seven classes into which these measurements might be grouped.

**(h1)Solution to Practice Problem 2**

Range of values is 168 – 36 = 32.

The approximate size of each class should be 32/7 = 4.54 ≈ 5.

The first class must include 136. Our last class must include 168. We could begin with 134 or 135 or 136 as our lower limit. We use the class width of 5 to establish the remaining lower limits. Once one upper limit has been determined, the remaining can be found using the class width of 5. Suitable answers would be:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | or |  | or |
| 134–138 |  | 135–139 |  | 136–140 |
| 139–143 |  | 140–144 |  | 141–145 |
| 144–148 |  | 145–149 |  | 146–150 |
| 149–153 |  | 150–154 |  | 151–155 |
| 154–158 |  | 155–159 |  | 156–160 |
| 159–163 |  | 160–164 |  | 161–165 |
| 164–168 |  | 165–169 |  | 166–170 |

**(h1)Practice Problem 3**

The number of drop-in (i.e., no appointment) patients per day at a community health clinic are grouped into a table with the classes 0–9, 10–19, 20–29, 30–39, and 40 or more. Will it be possible to determine from this table the number of days on which there were:

1. at least 20 drop-ins.
2. more than 20 drop-ins.
3. more than 19 drop-ins.
4. at least 19 drop-ins.
5. exactly 19 drop-ins.

**(h1)Solution to Practice Problem 3**

Classes 10–19, 20–29, 30–39, 40 or more.

1. Yes, by cumulating the frequency in the class 20–29 through to 40 or more (i.e., at least 20 means 20 or more).
2. More than 20 is not possible since there is no way to divide the class values 20–29 to exclude 20 and accept the other values.
3. More than 19 is equivalent to at least 20. See Part A.
4. At least 19 implies 19 or more. There is no way to divide the class values 10 to 19 to remove those values that are 19 and leave the other values in that class.
5. 19 alone cannot be isolated in the class 10–19. This is the problem with grouped versus raw data.

**(h1)Practice Problem 4**

The diastolic blood pressures of 120 male athletes between the ages of 15 and 20 were collected as part of a health survey conducted on athletes competing at the national level.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Diastolic Blood Pressure (mmHg)** | | | | | | | | | |
| 69  70  77  77  61  70  88  79  74  82  69  81 | 75  78  75  82  81  71  68  82  94  92  82  75 | 63  75  74  72  70  84  88  77  76  73  68  78 | 83  72  80  74  79  77  80  86  85  79  85  73 | 75  76  66  73  80  78  70  76  69  71  74  76 | 72  75  83  67  75  64  76  80  91  79  86  80 | 78  72  69  66  78  67  74  73  76  65  80  71 | 66  81  71  87  83  80  73  70  71  78  84  78 | 78  77  79  77  64  77  79  68  79  87  81  83 | 83  73  67  84  74  65  74  83  76  89  72  72 |

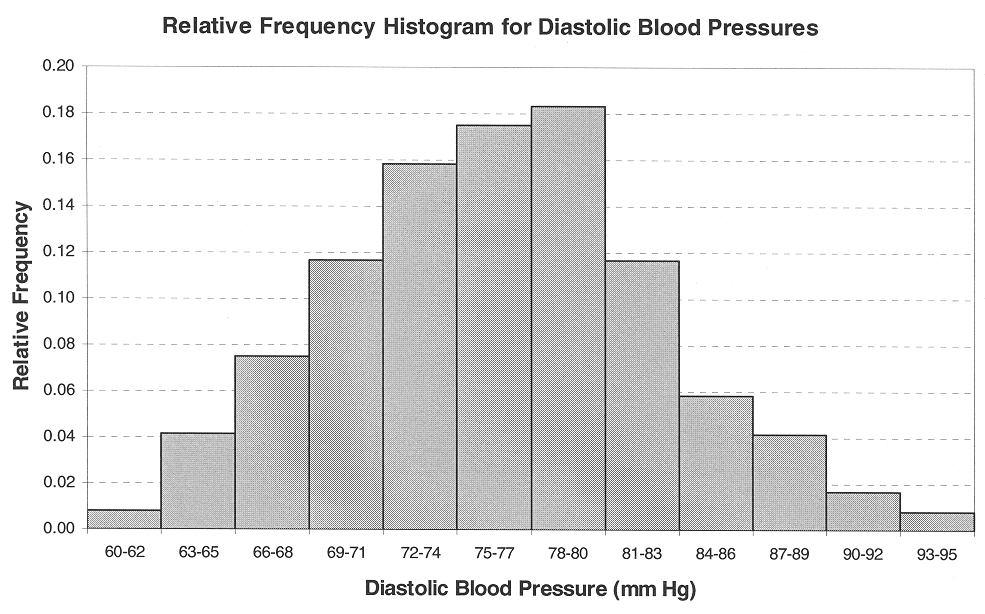
* 1. Describe the population of interest.
  2. Find the lowest and highest blood pressure in this table, and hence, find the maximum range in values.
  3. Give the class width if 10 intervals or classes were to be used. Choose the nearest whole integer for the class width. Divide the data into 12 equal classes using this width. Set up an appropriate frequency table, including headings of class number (1–12), class limits, class boundaries, tally, frequency, relative frequency.
  4. Construct a relative frequency histogram.
  5. What proportion of the blood pressures are less than 80.5 in the sample? What proportion of the area under the relative frequency histogram lies to the left of the 80.5? What is the relationship between the proportions? Can you give the proportion of blood pressures in the population which are less than 80.5?
  6. If you had chosen 10 or 14 class intervals, instead of 12, to span the range of the measurements, it is unlikely that the resulting frequency histogram would be identical to the one obtained in Part D. Is this important? Explain.
  7. How does the sample frequency histogram you drew relate to the unknown population frequency histogram?
  8. Construct a relative frequency polygon for the blood pressures.
  9. Construct a relative frequency ogive for the blood pressures.

**(h1)Solution to Practice Problem 4**

1. The population of interest is the set of diastolic blood pressures from all athletes competing at the national level from which only 120 athletes were sampled.
2. The lowest bp is 61 mmHg. The highest bp is 94 mmHg. The range of the 120 values is 94 – 61 = 33 mmHg.
3. If 10 intervals were to be used, the class width would be 33/10 = 3.3. The nearest whole integer is 3. The class width will be 3. As with Question 2, more than one answer would be acceptable. One such answer follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class i** | **Class**  **Limits** | **Class**  **Boundaries** | **Tally** | **Frequency**  **fi** | **Relative Frequency** |
| 1  2  3  4  5  6  7  8  9  10  11  12 | 60–62  63–65  66–68  69–71  72–74  75–77  78–80  81–83  84–86  87–89  90–92  93–95 | 59.5–62.5  62.5–65.5  65.6–68.5  68.5–71.5  71.5–74.5  74.5–77.5  77.5–80.5  80.5–83.5  83.5–86.5  86.5–89.5  89.5–92.5  92.5–95.5 | 1  ~~1111~~  ~~1111~~ 1111  ~~1111~~ ~~1111~~ 1111  ~~1111~~ ~~1111~~ ~~1111~~ 1111  ~~1111~~ ~~1111~~ ~~1111~~ ~~1111~~ 1  ~~1111~~ ~~1111~~ ~~1111~~ ~~1111~~ 11  ~~1111~~ ~~1111~~ 1111  ~~1111~~ 11  ~~1111~~  11  1 | 1  5  9  14  19  21  22  14  7  5  2  1 | 1/120  5/120  9/120  14/120  19/120  21/120  22/120  14/120  7/120  5/120  2/120  1/120 |

Normally you would include either class limits or class boundaries but not both. (***Note***: There is no ambiguity or overlap between the classes since each bp is measured to the nearest whole number.)



1. Proportion of sample blood pressures less than 80.5 is:



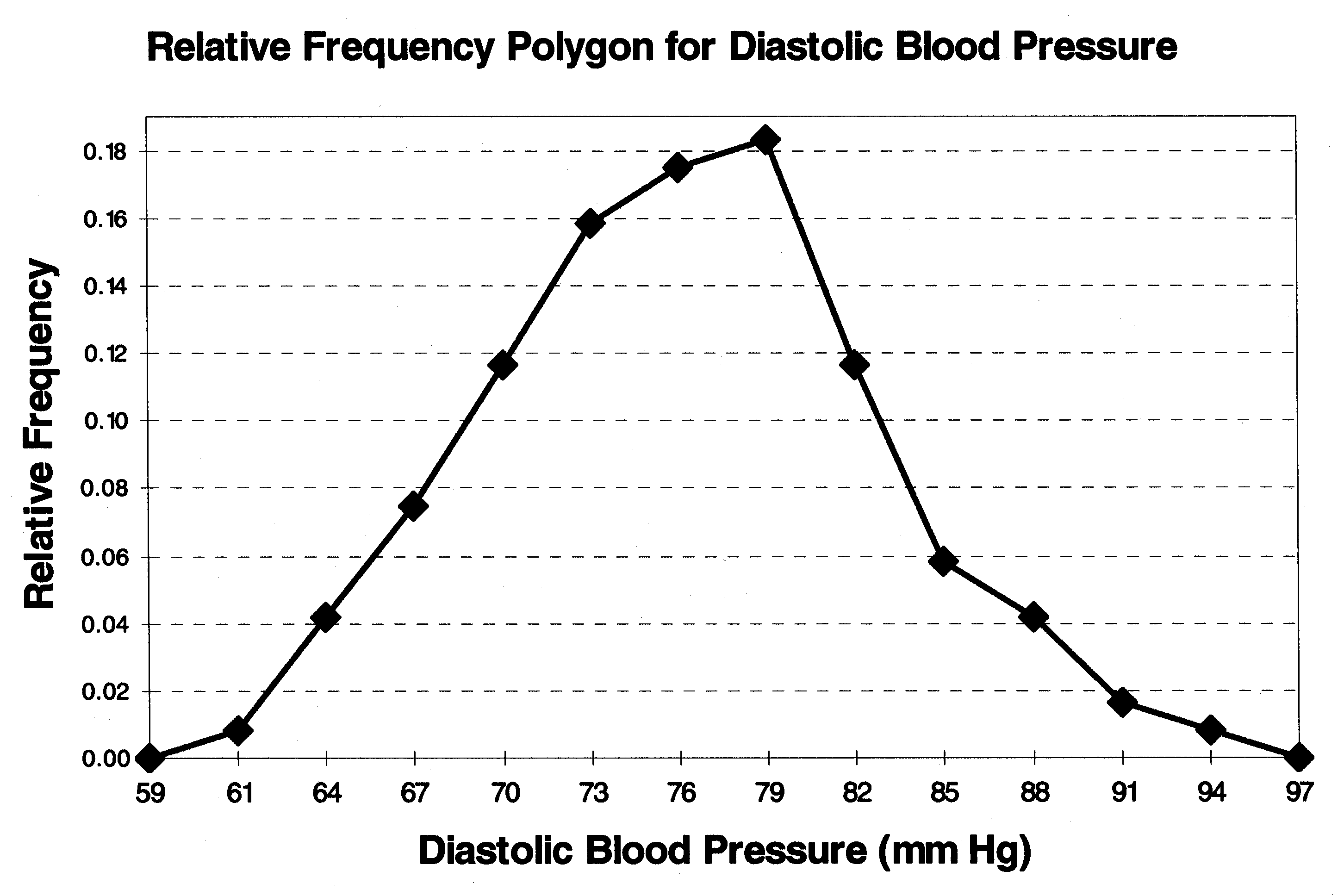
Proportion of area under the relative frequency histogram that lies to the left of 80.5 is also 91/120 (i.e., equals the sum of the area of the rectangles):



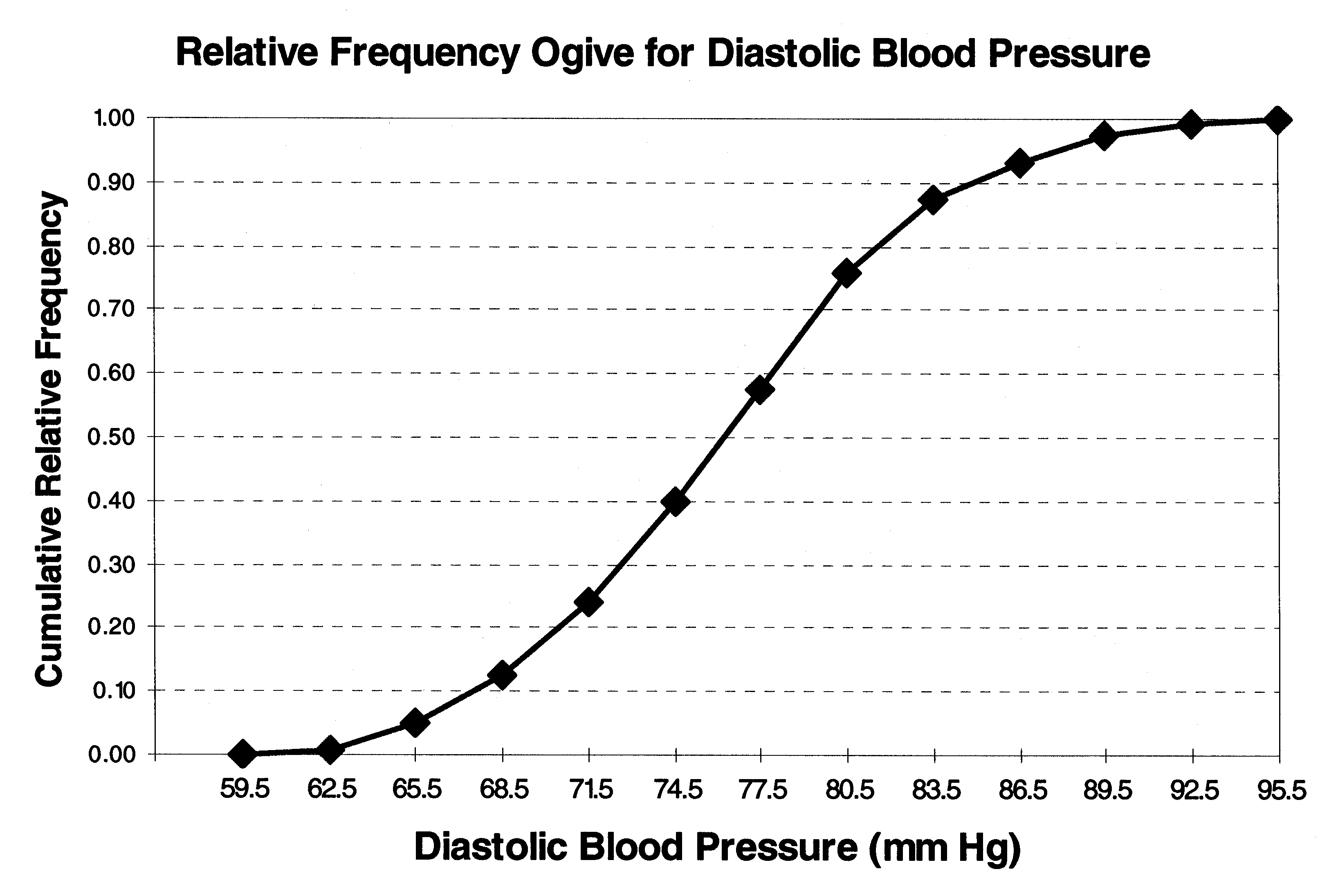
The proportions are equivalent.

Since it is likely that the sample histogram or sample frequency distribution is similar to the population frequency histogram or frequency distribution, we can infer that the proportion of blood pressures in the population that will be less than 80.5 is approximately 91/120. However, we need to place a bound on our error of estimation. We’ll learn to do this in later chapters.

1. To go from 12 to 10 or from 12 to 14 classes will not likely create large differences in the resulting histogram. However, it is important not to choose too few classes and risk obscuring some important observations or to choose too many classes and create a fairly meaningless frequency distribution with several empty classes. Generally, the more data, the more classes. As a rule of thumb, you should never have fewer than 5 or more than 20 classes regardless of the amount of data to be compacted. In practice, a computer would be used to group large amounts of data and you would try a few class sizes until you found the one which best depicted the visual picture of the underlying data.
2. It is likely that the sample frequency histogram, since it is based on 120 sample values, which is quite a large number of observations, is similar to the population frequency histogram.



1. .



**Module 2 Self-Test 1**

(mct) Some sort of test for data sets. Find pick the correct boundaries, pick the correct limits.

Confusing plot

**Numerical Summary of Data**

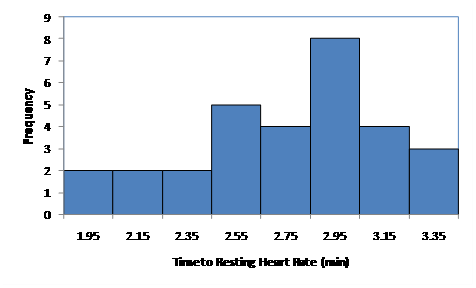
**(h1) Numerical Summary of Data**

**Calculation** **Note**: For the remainder of this course it will be expected that you will use a calculator to determine both the mean,, and the standard deviation, *s*, (or their population equivalents). This is a calculation that is done repeatedly throughout statistics so you should take this chance to determine how to get these values using your calculator. If you are uncertain and do not have your calculators manual, contact your instructor as they will have experience with many different types of calculators and can probably give you some help. Programs such as Excel™ can also be used for assignments but you will not have access to computers during your exams.

In the examples and practice problems, the calculation for mean and standard deviation is shown **but** this is not expected of you in your work.

**(h2) Measures of Central Tendencies**

The mode is most meaningful when there are a large number of data points and really with small data sets is really not that meaningful. For example, if we have some numbers {2, 4, 5, 6, 7, 7, 8} the mode is clearly and correctly identified as 7 but it is not clear why this is special since it only has one more occurrence than the other values. Additionally if your data is continuous you may have a region which occurs most frequently but no single data value that repeats. In this case, if we arrange our data into a histogram, we can call the class with the largest frequency, the *modal class* as it is the peak in our data. For example, looking at the histogram from an earlier example, we can see the modal class is the class which has a class mark of 2.95 minutes; this is the class for data values between 2.85 to 3.05 minutes and is fairly clearly a peak in our data.



Modal Class

Note that we will not consider weighted means although they may be used in cases such as calculating a student’s grade point average for a term or where some data may be considered more accurate than others and where a mean value is desired. However, this is not a procedure that should be used without justification and we will not make any weighted average calculations in this course.

**(h2)Measures of Variation**

In understanding standard deviation – in the research literature it is standard deviation not variance that is given as a measure of variation of data – it is important to realize that it can only be interpreted roughly in terms of how the data varies from the mean. It is almost but not quite the “average” deviation of the data from the mean value. However, when the data is symmetric and roughly “mound-shaped” then the Empirical Rule gives a clear relationship between where most of the data values are and what constitutes unusual data. Now it may seem strange to consider a special case of “mound shaped” but it happens that in real life data that consists of random variations about a mean value almost always follows such a distribution of values. This distribution is called a variety of names: Normal distribution, Gaussian distribution, or bell curve. For example, in measuring an electric current from a pacemaker, the current may be set to a specific value of 320 pA internally but in fact there are always small fluctuations about the set value if you measure carefully enough and these fluctuations being random will follow the Normal distribution. The following plot shows a histogram of some current measurements with an overlayed Normal distribution curve.

**(h2) Measures of Relative Standing**

The *z*–score is not used so much in the research literature in the context of summarizing data but as we will see in the (L to module 5) Hypothesis Testing Module (/L) it is used extensively with similar numbers such as a *t*-score to derive conclusions from research studies.

Be very careful with the quartiles (and  and all percentiles, as they are not rigidly defined mathematically; in principle they all have the same underlying definition but how it is interpreted differs. Different texts and programs calculate them differently and in this course the method used in the text will be adhered to. This means that in all likelihood if you use a calculator or computer program to determine these values, there is a good chance that your result will differ from the expected value.

**(h1) Example of Numerical Summary**

Using the data we saw in the last section for time to return to a resting heart rate,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Time to Return to Resting Heart Rate (min)** | | | | | |
| 1.9 | 2.4 | 2.6 | 2.9 | 2.9 | 3.2 |
| 2.0 | 2.5 | 2.7 | 2.9 | 3.0 | 3.2 |
| 2.2 | 2.6 | 2.7 | 2.9 | 3.0 | 3.4 |
| 2.2 | 2.6 | 2.7 | 2.9 | 3.1 | 3.4 |
| 2.4 | 2.6 | 2.8 | 2.9 | 3.2 | 3.4 |

1. Find the mean, median, mode.
2. Find the standard deviation using both your calculator and by estimating based on the Range Rule of Thumb (although generally don’t use this.)
3. Find the first and third quartiles
4. Find the 42nd percentile
5. Consider what happens to the all the above measures when the final data value is incorrectly entered as 34 rather than 3.4. Which values changed significantly and which stayed the same or similar.

**(h2) Discussion**

The first step is to arrange the data in ascending order

{1.9, 2, 2.2, 2.2, 2.4, 2.4, 2.5, 2.6, 2.6, 2.6, 2.6, 2.7, 2.7, 2.7, 2.8, 2.9, 2.9, 2.9, 2.9, 2.9, 2.9, 3, 3, 3.1, 3.2, 3.2, 3.2, 3.4, 3.4, 3.4}

1. Mean Median and Mode

To calculate the mean we can use our calculators or the equation given, noting that we have 30 data points.



The median is the same as the 50% percentile so we can use our percentile method to determine the median. First find the location in the list for the median

.

Being an integer this means that the median is halfway between the 15th data point and the 16th data point.



Finally for the mode we merely look through the list and find the number that occurs most frequently.

Mode= 2.9 as it occurs 6 times.

1. The sum of squares for this set of data gives so if we use the formula for the sample standard deviation



Using our calculator we get a number *s* = 0.396. The difference is solely due to rounding the mean value to 2.77 rather than keeping all its decimal places.

Using the Range Rule of Thumb

 a surprisingly good answer.

1. To locate the first quartile (Q1) we proceed in the same way as we did for the median.

.

Since this is not an integer we round up to the next integer and the first quartile must be the 8th data point in the list



To locate the third quartile (Q3)

.

Since this is not an integer we round up to the next integer and the first quartile must be the 23rd data point in the list



1. To locate the 42nd percentile (P42)

.

Since this is not an integer we round up to the next integer and P42 must be the 13th data point in the list



1. Repeating these calculations but with the last number changed to 34 rather than 3.4





Mode= 2.9



 no longer so good an answer.







Notice that the mean and the standard deviation both changed fairly significantly despite having only a single data point change. These measures are quite sensitive to the extreme data points in a set of data while the positional measures tend to be less sensitive to these extreme values.

Now try the following practice problems and compare your results with the solution provided.

**(h1)Practice Problem 1**

Consider the following problem.

Suppose a technologist is attempting to evaluate a new procedure for taking a computerized tomography (CT) scan of the chest. The quantity of interest is the time taken to complete the procedure. To compare the two approaches - new and old - for doing this measurement, she records the time (in minutes) for 25 CT scans of using each procedure done over a period of days.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Experimental Method** | | | | |  | **Standard Method** | | | | |
| 106  58  97  128  105 | 73  101  44  107  93 | 131  116  108  76  81 | 84  123  104  103  103 | 119  99  113  112  85 |  | 68  62  73  94  93 | 82  43  143  52  75 | 51  100  87  66  63 | 111  62  101  85  77 | 99  75  77  66  84 |

Compute the mean and standard deviation, s, for each set of data. What do these numbers tell you about the two groups of times? Compare the first, second and third quartiles as well. We will return to this problem in the next section and see how we can make this comparison graphical.

**(h1)Solution to Practice Problem 1**

***Experimental Method***:

Mean:  min.

Standard deviation:, giving



**Remember to sort the data before locating the quartiles.**

Q1: 

Since this is not an integer we round up to the next integer and the first quartile must be the 7th data point in the list 

Q2: .

Since this is not an integer we round up to the next integer and the second quartile (or median) must be the 13th data point in the list 

Q3: .

Since this is not an integer we round up to the next integer and the third quartile must be the 19th data point in the list 

***Standard Method***:

Mean: 

Standard deviation: 



Q1:  

Q2:  

Q3:  

The average time for the standard method is found to be ~ 19 minutes less than that of the experimental method with approximately the same standard deviation. Thus, in that the standard method provides a much better mean time with very little change in the “spread” of the data, it would appear that the standard method is more efficient (based on the given data and level of analysis). The quartiles tell us that almost all the standard method data values lie below the first quartile meaning that a significant fraction of the standard values are lower and the lower mean value is not just the result of some extreme values.

**(h1)Practice Problem 2**

The following data are a sample of cholesterol levels taken from 24 hospital employees who were on a standard North American (non-vegetarian) diet and who agreed to adopt a vegetarian diet for a 1-month period. Serum cholesterol measurements were made before adopting the diet and 1 month after.

|  |  |  |  |
| --- | --- | --- | --- |
| **Subject** | **Before** | **After** | **x = Before – After** |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24 | 195  145  205  159  244  166  250  236  192  224  238  197  169  158  151  197  180  222  168  168  167  161  178  137 | 146  156  178  146  208  147  202  215  184  208  206  169  182  127  149  178  161  187  176  145  154  153  137  125 | 49  11  27  13  36  19  48  21  8  16  32  28  13  31  2  19  19  35  8  23  13  8  41  12 |

1. Compute the mean change in cholesterol level.
2. Compute the appropriate standard deviation of the change in cholesterol levels.
3. Find the 60th percentile of the change in cholesterol levels.

**(h1)Solution to Practice Problem 2**

***Note***:  and , which greatly reduces our calculations but once again feel free to us your calculator.

Mean: 

Standard deviation:



To locate the 60th (P60)

.

P60 must be the 15th data point in the list of differences which must be first sorted



**Module 2 Self-Test 2**

Some sort of test for numerical measures. means, medians, skew quartiles.etc

**Exploratory Data Analysis**

**(h1)Exploratory Data Analysis**

Exploratory data analysis as was discussed in the text is the science (and sometimes art) of exploring data to look for trends or properties of the data. The boxplot is one tool which is particularly relevant to comparing data sets visually. In this course we will use modified boxplots as described in the text as these clearly label outlier data points and don’t give a false sense of the spread of the data which can occur with the skeletal boxplots. The following plot was taken from European Heart Journal (2006) **27**, pp 1571–1578 where they are using boxplots to compare myocardial blood flow (MBF) of various different categories of people in an effort to differentiate between left ventricular hypertrophy (LVH) in hypertensive patients and in athletes. The plots nicely demonstrate both the differences between the groups and also between the two situations of “at rest” and during induced hyperaemia.

Other powerful tools in EDA include the plots discussed earlier in this module. One restriction we will see is that if we do these procedures by hand it can be time consuming and thus we tend to not make several attempts using different class sizes. In research the use of a proper statistics program is almost mandatory to allow for easy data display and exploration.

**(h1)Example of Exploratory Data**

Using the data we saw in the last section for time to return to a resting heart rate,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Time to Return to Resting Heart Rate (min)** | | | | | |
| 1.9 | 2.4 | 2.6 | 2.9 | 2.9 | 3.2 |
| 2.0 | 2.5 | 2.7 | 2.9 | 3.0 | 3.2 |
| 2.2 | 2.6 | 2.7 | 2.9 | 3.0 | 3.4 |
| 2.2 | 2.6 | 2.7 | 2.9 | 3.1 | 3.4 |
| 2.4 | 2.6 | 2.8 | 2.9 | 3.2 | 3.4 |

1. Construct a boxplot.
2. Redo this boxplot with the last data point entered as 34 rather than 3.4.

**(h2) Discussion**

First we need our quartiles which we previously calculated









Any data point lower than  or larger than  would be considered an outlier.

The maximum data point that is not an outlier is 3.4 and the minimum data point that is not an outlier is 2.0 so the whiskers of the plot will go to these values. There is one outlier at 1.9.

b) Since all quartiles remain the same there are now 2 outliers, one at 1.9 and one at 34. The resulting boxplot is fairly useless as the box is shrunk by the existence of one very extreme outlier.

|  |  |
| --- | --- |
| a) | b) |

Now try the following practice problem and compare your results with the solution provided.

**(h1)Practice Problem**

Consider the following problem.

Suppose a technologist is attempting to evaluate a new procedure for taking a computerized tomography (CT) scan of the chest. The quantity of interest is the time taken to complete the procedure. To compare the two approaches - new and old - for doing this measurement, she records the time (in minutes) for 25 CT scans of using each procedure done over a period of days.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Experimental Method** | | | | |  | **Standard Method** | | | | |
| 106  58  97  128  105 | 73  101  44  107  93 | 131  116  108  76  81 | 84  123  104  103  103 | 119  99  113  112  85 |  | 68  62  73  94  93 | 82  43  143  52  75 | 51  100  87  66  63 | 111  62  101  85  77 | 99  75  77  66  84 |

Use two boxplots on the same scale to compare these two data sets.

**(h1)Solution to Practice Problem**

From previous practice problems we know:

***Experimental Method***:









Any data point lower than  or larger than  would be considered an outlier.

The maximum data point that is not an outlier is 131 and the minimum data point that is not an outlier is 58 so the whiskers of the plot will go to these values. There is one outlier at 44.

***Standard Method***:









Any data point lower than  or larger than  would be considered an outlier.

The maximum data point that is not an outlier is 111 and the minimum data point that is not an outlier is 43 so the whiskers of the plot will go to these values. There is an outlier at 143.

Two plots are shown. One used DataDesktopXL – ignore the shaded regions in the boxes) and the other used Minitab. Notice that there are some differences but the fundamental comparison between the two data sets is still the same. We can see graphically that the Standard method has most of its measurements below the median of the Experimental method.

|  |  |
| --- | --- |
| DataDeskXL Version | Minitab Version |

Note that if you did this plot by hand, the DataDeskXL version is the one to compare to.

**Module 2 Self-Test 3**

Compare some boxplots and some histograms of the same data

**Assignment 2**

Submit your assignment to the dropbox.