Grant's Tutoring

BASIC STATISTICS 1

Volume 1 of 2

September 2014 edition

Because the book is so large,
the entire Basic Statistics 1 course
has been split into two volumes.

Learn What You Need to Know
Know What You Need to Learn

While studying this book, why not hear Grant explain it to you? Contact Grant for info about purchasing **Grant's Audio Lectures**. Some concepts make better sense when you hear them explained.

Better still, see Grant explain the key concepts in person. Sign up for **Grant's Weekly Tutoring** or attend **Grant's Exam Prep Seminars.** Text or Grant (204) 489-2884 or go to **www.grantstutoring.com** to find out more about all of Grant's services. **Seminar Dates will be finalized no later than Sep. 25 for first term and Jan. 25 for second term.**

HOW TO USE THIS BOOK

I have broken the course up into lessons. Do note that the numbering of my lessons do not necessarily correspond to the numbering of the units in your course outline. Study each lesson until you can do all of my lecture problems from start to finish without any help. Then do the Practise Problems for that lesson. If you are able to solve all the Practise Problems I have given you, then you should have nothing to fear about your exams.

Although NOT ESSENTIAL, you may want to purchase the *Multiple-Choice Problems Set for Basic Statistical Analysis I (Stat 1000)* by Dr. Smiley Cheng. This book is now out of print, but copies may be available at The Book Store. The appendices of my book include complete step-by-step solutions for all the problems and exams in Cheng's book. Be sure to read the "Homework" section at the end of each lesson for important guidance on how to proceed in your studying.

You also need a good, but not expensive, scientific calculator. Any of the makes and models of calculators I discuss in Appendix A are adequate for this course. I give you more advice about calculators at the start of Lesson 1. **Appendix A in this book shows you how to use all major models of calculators.**

I have presented the course in what I consider to be the most logical order. Although my books are designed to follow the course syllabus, it is possible your prof will teach the course in a different order or omit a topic. It is also possible he/she will introduce a topic I do not cover. **Make sure you are attending your class regularly! Stay current with the material, and be aware of what topics are on your exam. Never forget, it is your prof that decides what will be on the exam, so pay attention.**

If you have any questions or difficulties while studying this book, or if you believe you have found a mistake, do not hesitate to contact me. My phone number and website are noted at the bottom of every page in this book. "Grant's Tutoring" is also in the phone book. **I welcome your input and questions.**

Wishing you much success,

Grant Skene

Owner of Grant's Tutoring and author of this book

Have you signed up for

Grant's Homework Help yet?

No? Then what are you waiting for? IT'S FREE!

Go to *www.grantstutoring.com* right now,

and click the link to sign up for

Grant's Homework Help

IT'S FREE!

- Grant will send you extra study tips and questions of interest throughout the term.
- You are also welcome to contact Grant with any questions you have. Your question may even provide the inspiration for other tips to send.
- If there are any changes in the course work or corrections to this book, you will be the first to know.
- You will also be alerted to upcoming exam prep seminars and other learning aids Grant offers.
- If you sign up, you will also receive a **coupon** towards Grant's services.

And, it is all FREE!

Four ways Grant can help you:

- **Grant's Study Books**

    - **Basic Statistics 1 (Stat 1000)**
    - **Basic Statistics 2 (Stat 2000)**
    - **Linear Algebra and Vector Geometry (Math 1300)**
    - **Matrices for Management (Math 1310)**
    - **Intro Calculus (Math 1500 or Math 1510)**
    - **Calculus for Management (Math 1520)**
    - **Calculus 2 (Math 1700 or 1710)**

    All these books are available at **UMSU Digital Copy Centre**, room 118 University Centre, University of Manitoba. **Grant's books can be purchased there all year round. You can also order a book from Grant directly.** Please allow one business day because the books are made-to-order.

- **Grant's One-Day Exam Prep Seminars**

    These are one-day, 12-hour marathons designed to explain and review all the key concepts in preparation for an upcoming midterm or final exam. Don't delay! Go to www.grantstutoring.com right now to see the date of the next seminar. A seminar is generally held one or two weeks before the exam, but don't risk missing it just because you didn't check the date well in advance. You can also reserve your place at the seminar online. You are not obligated to attend if you reserve a place. You only pay for the seminar if and when you arrive.

- **Grant's Weekly Tutoring Groups**

    This is for the student who wants extra motivation and help keeping on top of things throughout the course. Again, go to www.grantstutoring.com for more details on when the groups are and how they work.

- **Grant's Audio Lectures**

    For less than the cost of 2 hours of one-on-one tutoring, you can listen to over 40 hours of Grant teaching this book. Hear Grant work through examples, and offer that extra bit of explanation beyond the written word. Go to www.grantstutoring.com for more details.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

APPENDICES

SUMMARY OF KEY FORMULAS IN THIS COURSE

Lesson 1. sample standard deviation $= s = \sqrt{\dfrac{\sum\left(x_i - \bar{x}\right)^2}{n-1}}$

Lesson 2. correlation $= r = \dfrac{1}{n-1}\sum\left(\dfrac{x_i - \bar{x}}{s_x}\right)\left(\dfrac{y_i - \bar{y}}{s_y}\right) = \dfrac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\left(n-1\right)s_x s_y}$

slope $= b = r\dfrac{s_y}{s_x}$ intercept $= a = \bar{y} - b\bar{x}$

Lesson 4. standardizing formula for X bell curves: $z = \dfrac{x - \mu}{\sigma}$

Lesson 5. $P\left(A \text{ or } B\right) = P\left(A\right) + P\left(B\right) - P\left(A \text{ and } B\right)$

If A and B are independent: $P\left(A \text{ and } B\right) = P\left(A\right) \times P\left(B\right)$

Lesson 6. If X has a binomial distribution with parameters n and p, then the mean of $X = \mu_X = np$ and the standard deviation of $X = \sigma_X = \sqrt{np\left(1-p\right)}$.

The mean of $\hat{p} = \mu_{\hat{p}} = p$ and the standard deviation $= \sigma_{\hat{p}} = \sqrt{\dfrac{p\left(1-p\right)}{n}}$.

Also, $P\left(X = k\right) = \binom{n}{k} p^k \left(1-p\right)^{n-k}$.

Lesson 7. The mean of $\bar{x} = \mu_{\bar{x}} = \mu$ and the standard deviation $= \sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.

Central Limit Theorem: If n is large, $\bar{x}$ is approximately normal.

Standardizing formula for $\bar{x}$ bell curves: $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

Lesson 8. Confidence Intervals for μ: $\bar{x} \pm z^* \dfrac{\sigma}{\sqrt{n}}$ or $\bar{x} \pm t^* \dfrac{s}{\sqrt{n}}$

Sample size determination: $n = \left(\dfrac{z^* \sigma}{m}\right)^2$

Lesson 9. Test statistics for μ: $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ or $t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$

Lesson 11. Confidence interval for p: $\hat{p} \pm z^* \sqrt{\dfrac{\hat{p}\left(1-\hat{p}\right)}{n}}$

Sample size determination: $n = \left(\dfrac{z^*}{m}\right)^2 p^*\left(1-p^*\right)$

Test statistic for p: $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0\left(1-p_0\right)}{n}}}$

LESSON 1: DISPLAYING AND SUMMARIZING DATA

TABLE OF CONTENTS

First, it is essential that a statistics student have an adequate calculator. This does not mean an expensive calculator! You should be spending no more than $20. In my opinion, the **Sharp** scientific calculators are the best. They are easy-to-use, inexpensive, but have all the functions you would ever need. There will be a variety of model numbers, but the specific model you have is not important. Certainly, there are also **Casio** and **Texas Instrument** calculators that are good as well.

If you are going to purchase a new calculator, be it a Sharp, Casio, Texas Instruments, or whatever, here is a simple tip: **If the calculator is scientific and has four arrow buttons (left, right, up, down), then you can safely assume it does all the stats you would ever need, no matter what brand name it is.**

Many profs will recommend the Texas Instrument TI-30Xa. I disagree, this calculator is unable to do linear regression (see Lesson 2), and so is giving a student a definite disadvantage in assignments and exams. Expensive **graphing and programmable calculators** like the Texas Instrument TI-83 or similar, which many students used in high school, are also inappropriate because they **are specifically forbidden in your course outline**. I also consider them unnecessarily complicated.

Be sure to check Appendix A at the back of this book, for steps on how to use many makes and models of calculator. If you are unable to get your calculator to work after following my steps, do not hesitate to contact me by phone or through the "Homework Help" section on my website. **My phone number and web address are at the bottom of every page in my book.** I also encourage you to contact me with any questions you have about my book or the course in general.

Please go to www.grantstutoring.com right now and sign up for Grant's Homework Help. Every week I will send you extra homework help and tips including tips to help with the computer parts of this course. You will also be informed of any important course news. And, it's FREE.

TYPES OF VARIABLES

In Statistics, the entire group of individuals or objects we wish to examine is the **population**. The particular set of individuals or objects we select for study is the **sample**. Each individual in our sample is a **unit** (or, in the case of people, we can also call each individual a **subject**). **The number of units in our sample is the sample size and is denoted *n*.**

The characteristic of the unit we wish to measure is the **response variable** (or, simply, the variable), because the *response* can *vary* from one individual to another. We can ask a person a question and record their response. We can give a plant a certain amount of water (or none at all) and measure its response (how tall it grows perhaps).

We may select one sample, but measure several variables in that sample.* For example, we may be interested in the opinions of students enrolled full-time at the University of Manitoba, so we decide to survey 100 randomly selected students. We ask them several questions: How old are you? What is your major? How many hours do you spend on homework in a typical week? Do you own a computer? What is your GPA (Grade Point Average)?

In this example, the population we are examining is students enrolled full-time at the University of Manitoba; our sample is the randomly selected students we are surveying; our sample size is 100 (***n* = 100**); we are measuring five response variables in this sample: age, major, study time, computer ownership, and GPA.

We divide variables into two types: quantitative and categorical. **Quantitative variables are anything that can be measured numerically**, such as a person's height; the weight of a dog; the time it takes to drive to work. **Categorical variables** (which can also be called **qualitative variables**) **merely place each response into one of various categories** that are given or implied, such as a person's favourite colour; a student's major; the breed of dog found in Winnipeg.

* Do not confuse the word "sample" with "sample size". A sample is a group of objects or people we have selected for study. The sample size is how many objects or people we actually selected. When we say we have selected "one sample", that does not mean we have selected just one person or thing; we have selected a whole group of people or things. We are still waiting to hear what the sample size is, how many people or things have been selected.

Anytime a survey asks you a question and then asks you to select a category that best describes your opinion or situation, you are being asked for a response to a categorical variable. For example, if you were asked how much you spend on groceries in a typical week, that would be a quantitative variable (since you would give them a number like $50 or $100). On the other hand, if you were asked how much you spend on groceries in a typical week, and are given a choice of categories (less than $20, $20 to $100, over $100) that would now make the response a categorical variable.

Once we have collected data from a population or sample, we want to summarize the **distribution** in a clear and concise way. "Distribution" is a word you are going to see a lot in this course, so get used to it. Just like a corporation *distributes* bonuses to its employees (where some employees get much larger bonuses than others; usually in inverse proportion to their work ethic and morality), or newspapers *distribute* their product to various vendors and paperboys (where some people will get bigger piles of newspapers than others depending on the demand, and assuming anyone reads the newspaper anymore), it is the job of the statistician to determine the distribution of his/her data. What happens a lot? What happens rarely? We need to organize the data in such a way that anyone can see what the distribution is; which is to say, what the various responses are, and how frequent each response is.

QUANTITATIVE VARIABLES

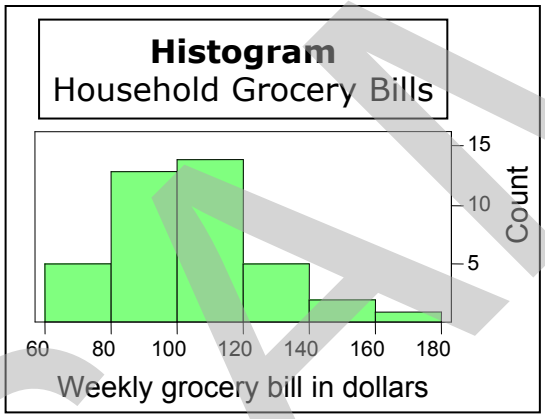
Quantitative variables allow us to do a lot of "number-crunching". At the very least, since the data is a bunch of numbers, we can line it up in ascending order (from lowest score to highest). We can then get a sense of which scores are high or low, which scores happen frequently, and which scores happen rarely. As we will see later, we can find means (averages) and medians, quartiles and standard deviations, and all sorts of other numbers that help us summarize the distribution.

We can subdivide quantitative variables into two types: **discrete** and **continuous**. **A discrete quantitative variable can have only certain numerical values and nothing in between.** For example, the number of children in a household is a discrete quantitative variable, because it could be 0 children, 1 child, 2 children, etc., but it could not be in between 1 and 2 children. You can't have 1.6 children or 2.4 children. (Pregnancy doesn't count; until that kid pops out of the womb, it's not getting a number!) Other examples of

discrete quantitative variables are the number you get when you roll a die (1, 2, 3, 4, 5, or 6, but, if you get anything in between when you roll (like 2.5), no one will want to play *Monopoly* with you); the price for different sizes of popcorn at a movie theatre ($2.99, $4.99, or $6.99, but you're paying nothing in between those prices, and please eat with your mouth closed).

A <u>continuous</u> quantitative variable can take on essentially any numerical value. For example, the height of 12-year old Canadian boys (a boy could be 44 inches tall and another could be 45 inches tall, but you could always find another boy, theoretically, whose height is in between those two boys). **Essentially, if you can conceive of a quantitative variable taking on "in between" values, then it is continuous.** Other examples of continuous quantitative variables are the *weight* of adolescents; the *amount of income tax* collected from Canadians annually; the *time* it takes to run 1500 metres.

The most common graph used to display the distribution of a <u>quantitative</u> variable (be it continuous or discrete) is a <u>histogram</u>. Typically, the <u>horizontal axis</u> shows the scores of the <u>response variable</u> we observed and the <u>vertical axis</u> shows the <u>frequency</u> of the scores (either by showing counts or percentages).



At left, is a histogram showing the weekly grocery bill of a random sample of 40 Canadian households. The horizontal axis shows the various bill amounts observed (*weekly grocery bill* is the response variable), and the vertical axis is the "count", how many households had the various bill amounts. At a glance, we have a good idea of the *distribution*. We see most households pay between $80 and $120 a week on groceries. (Measuring the bars according to the scale, it appears 12 households pay between $80 and $100, and 13 households pay between $100 and $120, making a total of 25 households out of the 40 paying between $80 and $120.) We can also see all the households paid between $60 and $180, but only 1 of them paid more than $160.

Two other graphs we can use to display quantitative variables are **stemplots** and **boxplots**. We will learn how to make and read those graphs later in this lesson.

CATEGORICAL (OR QUALITATIVE) VARIABLES

Categorical variables are much more limited in the kind of "number-crunching" we can do (that's why, in this course, we will generally be dealing with quantitative variables). All we can really compute is what percentage of the responses belongs to each category. For example, in a poll of likely voters in the next Federal election, we may have found 43% would vote Liberal, 20% Conservative, 15% Bloc Quebecois, 6% NDP, and 16% vote Other. But, that is about all we can do with this categorical variable.

We can subdivide categorical variables into two types: **ordinal** and **nominal**. **An ordinal categorical variable is where the proffered categories follow a logical order.** One end of the categories is clearly the opposite of the other end, with a neutral middle. For example, a survey asks your opinion on a proposal to construct an underpass to allow traffic to flow under a set of railway tracks at a cost of $40 million. You can choose one of the following responses. Are you strongly in favour, partly in favour, no opinion, partly against, strongly against? Here, there is a clear *order* to the categories ranging from people in favour to people against with neutrals sitting in the middle.

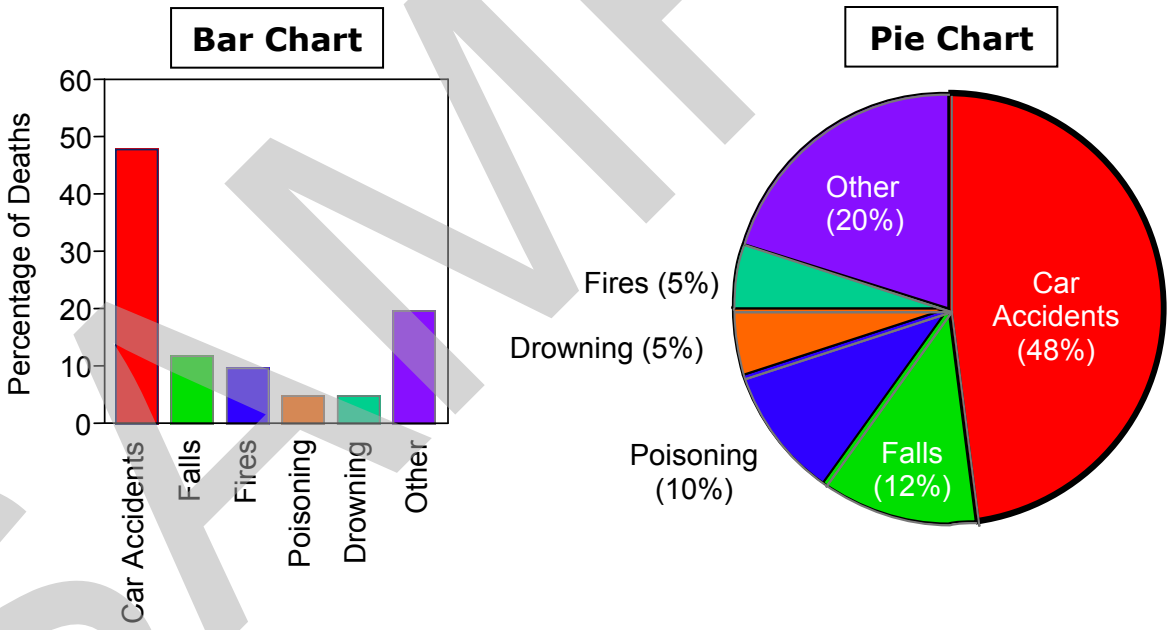
Often, ordinal categorical variables are coded numerically. In the example above, the survey could have asked for your opinion on this proposal on a scale of 1 to 5 where 5 means you are strongly in favour and 1 means you are strongly against the proposal. People will intuitively understand that 3 is neutral, while 4 is in favour, but not as strongly in favour as a 5.

Anytime, you are asked to rate something on a numerical scale, you are being presented with an ordinal categorical variable. Anytime you are given a choice of categories for your response, and the categories have a clear progression from one extreme to another, you are being presented with an ordinal categorical variable. For example, instead of simply being asked what is your household's weekly grocery bill (where the response variable is quantitative), you are asked to choose one of these categories: under $25, $25 to $75, $75 to $150, over $150. These categories have a clear progression ranging from "not very much" to "a heck of a lot", making *weekly grocery bill* an ordinal categorical variable. **A researcher always has the option of expressing a quantitative variable (be it continuous or discrete) as an ordinal**

categorical variable instead if that suits their purposes better. It depends on how exact they need their numerical data to be.

A nominal variable is where the categories (given or implied) simply have names. The order of the categories is arbitrary, often simply alphabetical. For example, we may ask U of M students do they belong to the Faculty of Arts, Commerce, Engineering, Nursing, Science, or none of the above? We could list the faculties in any order we want (obviously, "none of the above" would be at the end of the list).

If we wish to graph our results for a categorical variable, we can use a **bar chart** or a **pie chart**. Below are examples of a bar chart and pie chart for the various causes of accidental death in the United States. Note that we do not have to put the categories in any particular order (like most common to least common, for example).



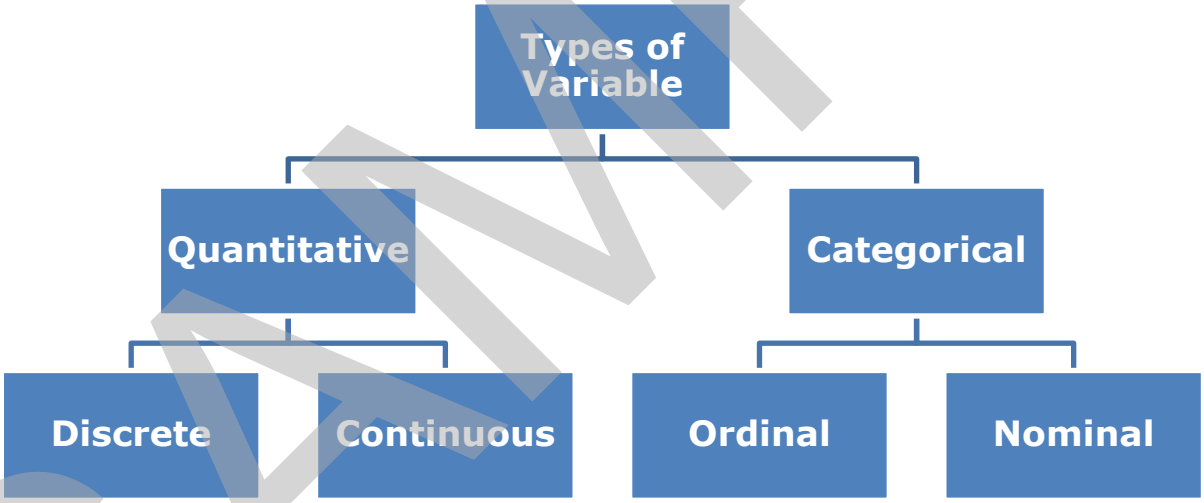
HISTOGRAMS VS. BAR CHARTS

Don't confuse a histogram with a bar chart! They do look similar. The difference is in the horizontal axis. A histogram has a numerical scale on its horizontal axis because it is displaying the distribution of a quantitative variable. A bar chart has the various names for its categories labelled on the horizontal axis because it is displaying the distribution of a categorical variable.

Simply put:

Histograms are for quantitative variables.

Bar Charts are for categorical variables.



Display quantitative variables with histograms, stemplots or boxplots.*

Display categorical variables with bar charts or pie charts.

Let's try some questions.

* I know! You don't even know what a stemplot or a boxplot is yet! Nonetheless, these graphs, as well as a histogram, can be used to display any quantitative variable. You will see what stemplots and boxplots are, and how to make them later in this lesson.

1. A survey asked the following questions:

- **What is your eye colour?**
- **Rate your boss on a scale from 1 to 10 where 1 is awful and 10 is wonderful.**
- **How much do you weigh (in kilograms)?**
- **What is the 3-digit area code for your home phone number?**
- **How far do you live from work or school? (less than 5 km, between 5 km and 10 km, more than 10 km)**
- **What method of transportation to work/school do you normally use? (car, bus, bike, walking, other)**
- **What is your annual income?**
- **How many times in a typical month do you eat at a restaurant (including take-out or delivery)?**
- **Did you vote in the last federal election?**

For each of the above survey questions, identify if the variable is quantitative (if so, is it discrete or continuous?) or categorical (if so, is it ordinal or nominal?). In addition, what graph could you use to display the data gathered in each case?

If you ever get a big, long question like this on an exam or assignment (and you will), never read the whole thing! Skip to the last sentence or two and read that first. That will probably tell you what the question really wants you to do. Once you know what your goal is, you can go back to the start of the question and skim through to see what you need.

In this case, the last two sentences make it clear we have to determine if each variable in the survey is quantitative (discrete or continuous) or categorical (ordinal or nominal). We also have to tell them what graph we would use to display the data in each case.

As far as displaying the data is concerned, we learned above histograms, stemplots or boxplots can be used to display quantitative variables (both continuous and discrete), and bar charts or pie charts can be used to display categorical variables (both ordinal and nominal). So,

the minute we know what type of variable it is, we know what graphs we can use.

Let's take each variable in turn:

- ***Eye colour* is a nominal categorical variable. We could display the data with a bar chart or a pie chart.** We are choosing categories (blue, green, brown, etc.) which have no special order so it is merely nominal, not ordinal.
- ***Boss rating* is an ordinal categorical variable. We could display the data with a bar chart or a pie chart.** As we saw earlier, any variable that asks you to rate something on a numerical scale is an ordinal categorical variable. This ten-point scale is simply giving us ten categories to choose. Obviously, the categories have a logical *order* taking us from one extreme (awful) to another (wonderful), making the variable *ordinal*.

Is that variable quantitative or not?

Not all numerical variables are quantitative. We already know variables that ask you to rate something on a numerical scale are ordinal categorical variables. Other kinds of numerical variables are not quantitative either.

If you ever find yourself unsure whether a variable is quantitative or not, ask yourself this question: "Would an average be meaningful in this context? Would it make sense to call a score *above* average or *below* average?" If the answer is "yes", then you must have a quantitative variable. If the answer is "no", then it is not (it must be a categorical variable).

- ***Weight* is a continuous quantitative variable. We could display the data with a histogram, stemplot or boxplot.** Obviously, we could stand on a scale and get a clear numerical measurement of our weight. Someone could weigh 50 kilograms, someone else 51 kilograms, and someone else could weigh anything in between, so it is continuous not discrete. It makes perfect sense to talk about the average weight and someone whose weight is below or above average.

- ***Area code* is a categorical and nominal variable. We could display the data with a bar chart or a pie chart.** Even though it is a number, it is not quantitative. The average of all the area codes would be meaningless. I have a below average area code. Huh? Nobody cares if your area code is high or low; above average or below average. Being asked for your area code is like being asked what city do you live in. Just as *city* is categorical and nominal (cities could be listed in any order you please, so they are not ordinal), so is *area code*. Certainly, you could line the area codes up from lowest number to highest, but that is not a meaningful order. That is no different from lining categories up alphabetically. The order must be more meaningful and necessary than that for a categorical variable to be ordinal.
- ***Distance from work/school* is an ordinal categorical variable. We could display the data with a bar chart or a pie chart.** If they had simply asked us for a distance, it would have been quantitative and continuous, but *we are forced to select a category*, making it a categorical variable. There is a logical progression from closest to farthest away in the categories, thus ordinal.
- ***Method of transportation* is a nominal categorical variable. We could display the data with a bar chart or a pie chart.** We have been given categories to choose which could have been given in pretty much any order, so it is certainly not an ordinal variable.
- ***Annual income* is a continuous quantitative variable. We could display the data with a histogram, stemplot or boxplot.** We would give them a number (quantitative) which could be, for example, $200, $201 or anything in between (continuous). Some people will make below average incomes and others will be above average.
- ***Restaurant visits* is a discrete quantitative variable. We could display the data with a histogram, stemplot or boxplot.** Some people could eat restaurant food 0 times a month, 1 time, 2 times, etc., but you can't eat at a restaurant an "in between" amount such as 1.3 times in a month. That makes *restaurant visits* a *discrete* quantitative variable. As with all quantitative variables, it is reasonable to talk about the average number of restaurant visits. In fact, the average number of visits could be an

impossible amount like 3.7 times a month. That still has meaning. We would now know someone who visits a restaurant 5 times a month is above average, for example.

Even though discrete quantitative variables cannot have "in between" values, there is nothing wrong with the average value being one of those "in between" values. (The average family has 2.2 children even though no actual family could.) Averages merely give us a gauge to tell which scores are below average and which are above.

▶ ***Voting history* is a nominal categorical variable. We could display the data with a bar chart or a pie chart.** Any yes/no, male/female kind of question where you can only choose one of two categories, even though they are opposites, should always be considered nominal, not ordinal.

Solution to Question 1

***Eye colour* is a nominal categorical variable. We could display the data with a bar chart or a pie chart. *Boss rating* is an ordinal categorical variable. We could display the data with a bar chart or a pie chart. *Weight* is a continuous quantitative variable. We could display the data with a histogram, stemplot or boxplot. *Area code* is a categorical and nominal variable. We could display the data with a bar chart or a pie chart. *Distance from work/school* is an ordinal categorical variable. We could display the data with a bar chart or a pie chart. *Method of transportation* is a nominal categorical variable. We could display the data with a bar chart or a pie chart. *Annual income* is a continuous quantitative variable. We could display the data with a histogram, stemplot or boxplot. *Restaurant visits* is a discrete quantitative variable. We could display the data with a histogram, stemplot or boxplot. *Voting history* is a nominal categorical variable. We could display the data with a bar chart or a pie chart.**

Rule of thumb: if only two categories are offered, and there is no possible way there could have been more, then we have a nominal variable for sure (yes/no; true/false; agree/disagree). If you are given only two categories but could conceivably have more to choose from, and, if those categories would have a logical order, then you have an ordinal variable. For example, "How many children do you have? Two or less? More than two?" gives us only two categories to choose from, but could have easily had more categories in a nice increasing order, making it a categorical and ordinal variable.

2. You record the age, marital status, earned income, and sex of a sample of 1463 people. The number of variables you have recorded is:

(A) 1463.

(B) Five: age, marital status, income, sex, and number of people.

(C) Four: age, marital status, income, and sex.

(D) Two: age and income; marital status and sex are not variables because they are not a numerical quantity.

(E) None: because no one has any business asking such personal questions.

There are four variables: 2 quantitative variables (age and income) and 2 categorical variables (marital status and sex). (Choice (D) is wrong because it thinks only the quantitative variables should count.) These four items are *variables* because each one of them has a *variety* of answers. 1463 is not the number of variables; it is the size of the sample (n = 1463).

Solution to Question 2

The correct answer is (C).

3. The table below shows the number of people (in millions) living on farms in a certain country over the years.

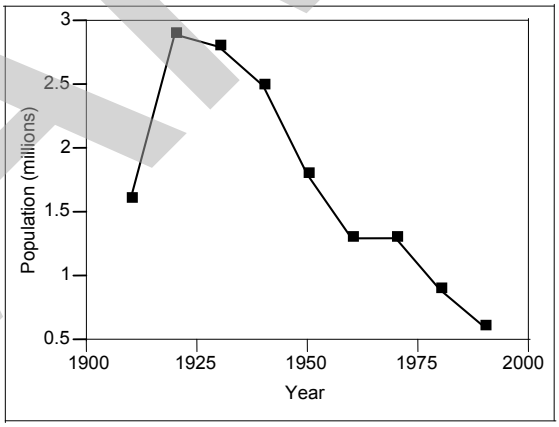
Year	1910	1920	1930	1940	1950	1960	1970	1980	1990
Population	1.6	2.9	2.8	2.5	1.8	1.3	1.3	0.9	0.6

Construct a time series for this data and comment on what you see.

Frequently, a researcher wants to see how a quantitative variable changes as time goes by. They measure the variable at regular time intervals and determine if there is a trend. To help see any trends, they construct a **time series** (also called a time plot). **Use a time series to display any data that has been collected as time goes by.**

A time series is a very simple graph to draw. **The time unit is always put on the horizontal scale, and the other quantity we are measuring is put on the vertical scale. Make sure you label your axes and title the graph!** We then plot dots for each observation and connect the dots. Having done that, we then look for any trends: Are the observations falling as time goes by, or rising, or a bit of both? Did the trend change at any time?

Solution to Question 3



Apart from a sudden jump in the first decade (1910 to 1920), we see a steady downward trend in farm population. It appears the number of people living on farms is declining in this country.

4. The test scores (out of 100) for a random sample of 50 students who wrote a statistics midterm exam are as follows:

75	88	47	66	78	45	73	66	77	100
64	61	77	87	66	92	86	57	80	70
52	84	80	79	66	92	72	83	50	84
65	75	77	79	79	57	63	51	44	59
84	77	44	81	61	77	57	75	3	52

(a) Construct a frequency table.

(b) Construct a relative frequency table.

(c) Construct a histogram.

(d) Construct a stemplot.

(e) Construct a split stemplot.

(f) Discuss the shape of the distribution. Are there any outliers?

(g) Find the median test score.

(h) Find the first and third quartiles.

(i) Find the Range and Interquartile Range.

(j) State the five-number summary.

(k) Justify the outliers (if any) mathematically.

(l) Draw a boxplot.

(m) Draw a modified (outlier) boxplot.

(n) Find the mode of the distribution.

(o) Find the mean of the distribution.

(p) Find the standard deviation of the distribution.

In this problem, we have a random sample of 50 statistics students (***n* = 50**). The variable is ***test score***, a **continuous quantitative variable**. You might think the variable is discrete since all fifty test scores are nice whole numbers (no decimals). It doesn't matter what the data is, but what it might have been. One student may score 77 on the test and another score 78, but, conceivably, a third could score something in between (assuming the students could get part marks on questions). Admittedly, if it is impossible to get a mark between 77 and 78 (because there are 100 questions that are either right or wrong, for example), then *test score* would be a discrete quantitative variable. Be the variable discrete or

continuous, the important thing is it is quantitative, and all quantitative variables are open to the same analysis.

This question illustrates all of the kinds of "number-crunching" a statistician can do to visualize and describe the distribution of a quantitative variable. A well-presented summary of a distribution should answer the questions, "What happened, and how often?" **Specifically, when summarizing the distribution of a quantitative variable, the three things we should cover are the shape, centre and spread.**

To help make sense of this sample, first we need to arrange it into some meaningful order, and that is what part (a) is having us do.

4. (a) Construct a frequency table.

First of all, do not think for a minute you will ever have to do something like this on an exam! It is far too time-consuming to make a frequency table. As long as you can read and interpret a frequency table that they give you, you know all you need to know. I include this question only because you may be asked to do something like this on an assignment (but even that is unlikely). Other than that, you can safely assume you will never have to make a frequency table again! Another reason that this would never be a test question is there is no such thing as only one correct answer for a question like this. Many different arrangements are possible for any given set of data, although some may do a better job of summarizing the data than others.

As the name implies, a frequency table organizes the data in such a way that we can see how *frequently* various values came up. To make it easier to comprehend, we break the data into **classes** of equal size. That means we must first decide on our **class boundaries**. This is where most students panic. They wonder, "How do I know what to use for class boundaries?" Don't worry about crap like that! They will tell you what to do (they have to in order to ensure there is only one correct answer). If they don't, they have just forced themselves to accept pretty well anything as correct.

If you have to come up with your own classes, keep it simple, counting by tens or fives. The goal is not to use too many classes or too few. A helpful guide is "**the $\sqrt{n}$ rule**" (where n is the sample size). Compute $\sqrt{n}$ and round off to give you an idea of how many classes to use.

This is only a guide, and you can use slightly more or less classes if you feel that will keep the numbers neat.

In our problem, $n = 50$; therefore, $\sqrt{n} = \sqrt{50} = 7.07\ldots$, so something like 7 classes would be a good choice. Looking at the data, we see the minimum value is 3 and the maximum is 100, so, if we started at 0 and went by tens, we would end up with 10 classes. This seems to be a little more classes than we would like, but who cares? Certainly, not me! Don't obsess about junk like this! They didn't tell us how many classes to make, so I figure 10 classes is close enough to my target of 7 to satisfy me.

This number of classes is really due to the minimum value of 3. The next lowest value is 44, meaning we would have been able to start our classes at 40 if it had not been for this one unusually small value. Consequently, it seems perfectly fine to go by tens since 49 out of the 50 data values would then end up in 6 classes. The number 3 is an example of an **outlier** (a data value falling outside the overall pattern of the data).

Another thing to be aware of is to **ensure there is no confusion as to which numbers belong to which class**. If the data are all whole numbers (no decimals), as in this case, we can accomplish this by simply using classes with no overlapping numbers. i.e. We could use as our classes: 0 – 9, 10 – 19, 20 – 29, … 100 – 109; or, better yet, 1 – 10, 11 – 20, 21 – 30, … 91 – 100, perfectly suiting the fact that 100 is the highest possible score. **It is important to understand that there is no wrong answer when you have to choose the classes yourself.** As long as you have chosen a reasonable number of classes, and there is no confusion which numbers belong to which class, your particular choice of class limits cannot be considered wrong.

Generally, it is better to present **continuous** classes. This is where the boundaries join together, the upper limit of a class boundary is the lower limit of the next class boundary. i.e. We could use as our classes: 0 to 10, 10 to 20, 20 to 30, … 90 to 100 (which also could be written with a "dash" if you prefer: 0 – 10, 10 – 20, 20 – 30, … 90 – 100). But, the problem here is we are left unsure where data belongs. For example, if someone scored 30 on the test, would they be in the 20 – 30 class or the 30 – 40 class? Even if someone did not score 30 on the test (as is the case in our set of data), it is still considered unacceptable to have class limits that are potentially ambiguous like this.

This is easily rectified by simply making it clear to people which of the classes would contain 30. You could write a note telling people the left endpoint is included but the right endpoint is not (i.e. for 10 – 20, a score of 10 would be considered part of this class (since 10 is the left endpoint), but a score of 20 would not be in this class (since 20 is the right endpoint), 20 would be included in the 20 – 30 class instead). Alternatively, you could say the reverse: the left endpoint is not included, but the right endpoint is for each class. That would be preferable for this data since that would mean 90 – 100 would include 100 now (but not include 90).

Rather than write a note explaining this, we can make it clear in our notation which endpoints are included and which are not. For example, to say our variable, x, is between 0 and 10, we literally write x between 0 and 10 and use "less than" signs ($<$). $0 < x < 10$ literally says 0 is less than x which is less than 10, but we simply read this as x is between 0 and 10. However, that means neither endpoint is included. To include an endpoint in the class, we use a "less than or equal to" sign ($\leq$) instead. If we wish to include the left endpoint but not the right, we would write $0 \leq x < 10$, $10 \leq x < 20$, $20 \leq x < 30$, etc. Similarly, if we want to include the right endpoint instead of the left, we would write $0 < x \leq 10$, $10 < x \leq 20$, $20 < x \leq 30$, etc. Properly, we should also tell the reader what x stands for. Simply define x as the variable you are measuring in the problem. Here, we would say let x = the test score.

A less cumbersome way of accomplishing this would be to use interval notation. This is a standard mathematical method to show a range of values. In this notation, we enclose the classes in brackets; a square bracket, "[" or "]", tells you to include the endpoint, a round bracket, "(" or ")", tells you do not include the endpoint. In this notation, the endpoints are separated by a comma, not a dash. For example, [10, 20) includes a test score of 10 but does not include 20; whereas, (10, 20] includes 20 in its class but does not include 10. Thus we could use the class limits (0, 10], (10, 20], (20, 30], etc.

Yet another way to define your classes to ensure no confusion is to use one more decimal place in the classes than the data has itself. This way you can allow the class limits to overlap. For this question, we could use the class limits of 0.5 – 10.5, 10.5 – 20.5, 20.5 – 30.5, ... 90.5 – 100.5. If our data had one decimal place in it (e.g., 41.3, 52.5, etc.), we would use 2 decimal places in our class limits such as 0.05 – 10.05, 10.05 – 20.05, 20.05 – 30.05, ... 90.05 – 100.05, but that is getting pretty silly! Note: the last decimal place in the class limits should always use a "5" for its digit to ensure no confusion.

In general, if using the same number of digits in your classes as the data has itself, make sure there is no confusion in your class limits, either by having no overlap (e.g. 1 to 10, 11 to 20, 21 to 30, etc.) or by making it clear which endpoint is included and which is not (e.g. $0 < x \le 10$, $10 < x \le 20$, $20 < x \le 30$, etc.). If using one more decimal digit in your class limits be sure the last digit is a "5", and then the upper limit for one class acts as the lower limit for the next class with no possible confusion (e.g. 0.5 to 10.5, 10.5 to 20.5, 20.5 to 30.5, etc.).

Again, none of this is really important. You will probably never be stuck having to make a decision what class limits to use. I have done this just to show you all the different ways you may see a frequency table presented in class or on an exam. If you have to make your own table with no direction whatsoever, pick whichever of the methods above you like and use it.

I recommend you always use some type of continuous set of class boundaries where one endpoint is included but the other is not because it is simple to read and avoids making numbers any more complicated. It also sets you up for making a histogram (as we are asked to do in part (c) of this question).

For this problem, I am going to use 0 – 10, 10 – 20, 20 – 30, etc. for my classes. This style is very easy to understand for anyone reading your table, and does not get into the complications of introducing decimals. I will include a note telling people to exclude the left endpoint in each class but include the right endpoint. Generally, we prefer to include the left endpoint rather than the right endpoint, but that is not very suitable here since that would necessitate a 100 – 110 class just to be able to include the score of 100.

Once you have decided on your classes, you tally up all the data to see how much fits in each class. I simply go through the list of data, ticking off which class it fits. For example, reading the data from left to right, the first test score is 75, so I put a tally mark in the 70 – 80 class; the next score is 88, so I put a tally mark in the 80 – 90 class; the next score is 47, so I put a tally mark in the 40 – 50 class; etc..

Note, since I have decided to exclude the left endpoint in each class and include the right endpoint, the person who scored 80 is given a tally mark in the 70 – 80 class while the person who scored 70 is given a tally mark in the 60 – 70 class.

When you have completed your tally, check that it totals up to the correct amount. Here, we know $n = 50$, so we better have a total of 50 tally marks.

Test Score*	Tally
0 – 10	𝍸
10 – 20	
20 – 30	
30 – 40	
40 – 50	𝍸
50 – 60	𝍸 II
60 – 70	𝍸 𝍸
70 – 80	𝍸 𝍸 𝍸 I
80 – 90	𝍸 III
90 – 100	III

* Each class includes the right endpoint, but excludes the left endpoint. e.g. "70 – 80" excludes test scores of 70, but includes test scores of 80.

We can now present our Frequency Table. (**Note, the results of our tally are recorded in a column called "Frequency" or "Count", while we use our variable to name the column of classes, here "test score".**) Be sure to include a note to explain which endpoint is included in each class.

Solution to Question 4(a)

Test Score*	Frequency (or Count)
0 – 10	1
10 – 20	0
20 – 30	0
30 – 40	0
40 – 50	5
50 – 60	7
60 – 70	10
70 – 80	16
80 – 90	8
90 – 100	3

*** Each class includes the right endpoint, but excludes the left endpoint. e.g. "70 – 80" excludes test scores of 70, but includes test scores of 80.**

We still include the classes "10 – 20", "20 – 30, and "30 – 40" even though there is no data in those regions. We simply put "0" in their count column.* This helps a reader see the one score in the "0 – 10" is unusually low. We would say "0 – 10" is an **outlying class** since it is so much lower than all the other scores. Specifically, if we had not noticed before now, we can tell the person who scored 3 on the test is an outlier.

The point to making a frequency table is giving us a handle on what scores are common and what scores are rare (or nonexistent). For example, we have now discovered all but one of the fifty students scored above 40 on the test; only six students scored 50 or less; more than half of the students scored between 60 and 80 (10 in the "60 – 70" class plus 16 in the "70 – 80" class equals 26 out of 50 students). There are, of course, many other things we could observe.

* By the way, I labelled the second column "Frequency (or Count)". You should not do that; it is redundant. You can label the column "Frequency" or label it "Count", which ever you prefer. I just want you to be aware of the different labels that could be used.

4. (b) Construct a relative frequency table.

To make a relative frequency table, simply **divide each frequency value by *n***, and convert the decimal to a percent value (it is also fine to leave the results in decimal form, although percent is more typical). **Remember, to change a decimal into a percent, multiply it by 100%, or simply move the decimal two places to the right.** Here, $n = 50$ so, 1/50 = .02 or 2%, 5/50 = .10 or 10%, 7/50 = .14 or 14%, … 3/50 = .06 or 6%.

Solution to Question 4(b)

Test Score*	Relative Frequency (or Percentage)
0 – 10	2%
10 – 20	0%
20 – 30	0%
30 – 40	0%
40 – 50	10%
50 – 60	14%
60 – 70	20%
70 – 80	32%
80 – 90	16%
90 – 100	6%

*** Each class includes the right endpoint, but excludes the left endpoint. e.g. "70 – 80" excludes test scores of 70, but includes test scores of 80.**

The "Relative Frequency" column should total up to 100%. Sometimes, you may get very messy results that require rounding off. For example, maybe $n = 30$ and you have a frequency of 7, then you would get 7/30 = .23333… or 23.3% (expressing your percentage rounded off to one decimal place would be a good rule of thumb; do not round off to the nearest whole number!). When rounding off was necessary, your "Relative Frequency" column may not total to exactly 100%, but it should be very close (like 99.9% or 100.1%, no worse).

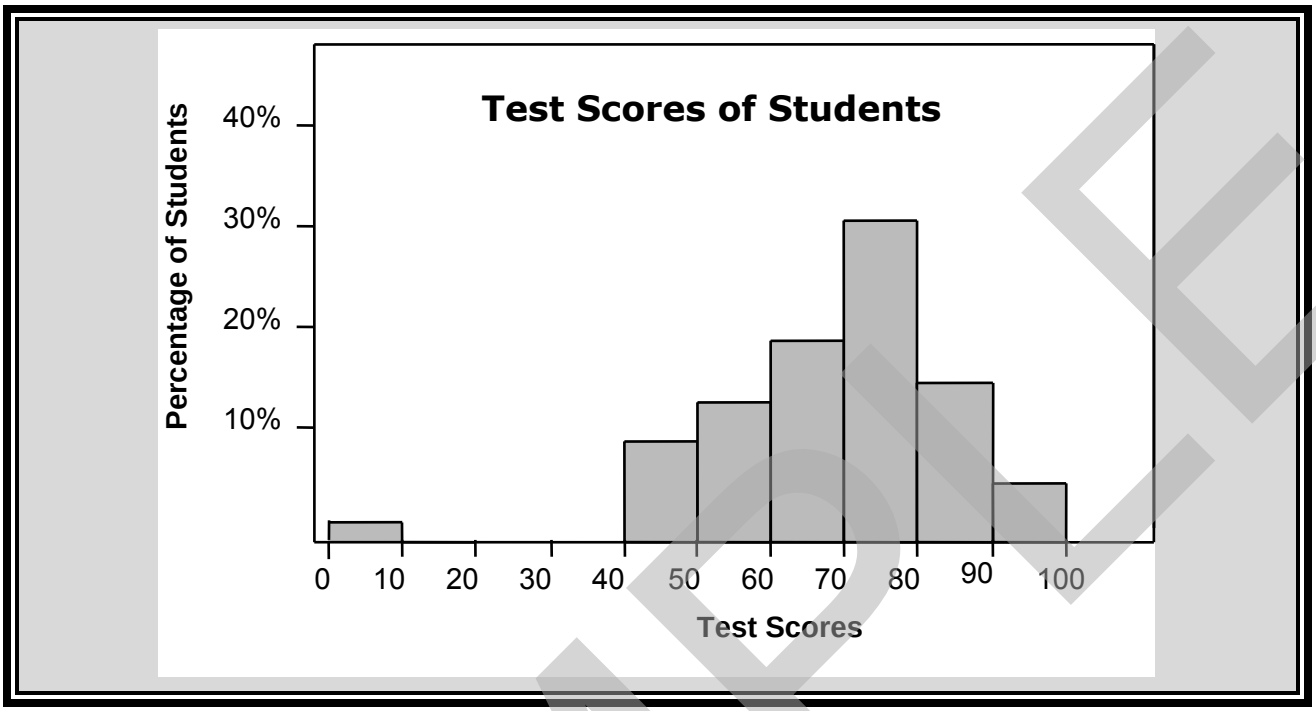
4. (c) Construct a histogram.

A histogram is merely a picture of our frequency or relative frequency tables. The **class limits** will give us the scale on the **horizontal axis** and the **frequency** (or **relative frequency**, whichever you prefer if not specified) will be marked on the **vertical axis**. We then draw bars to represent the frequency in each class. **Do not forget to label the axes!**

Since we have a <u>continuous</u> quantitative variable, the bars on the histogram for adjoining classes cannot have any gaps between them. Gaps between the bars of a histogram imply the variable cannot have any "in between" values. Which is to say, gaps imply the quantitative variable is discrete. If you constructed a frequency table with gaps in your class limits, then you would have to redefine them to get rid of the gaps. For example, if you had used classes of 41 to 50, 51 to 60, etc., leaving a gap between 50 and 51, we cannot mark both 50 and 51 on the horizontal axis, draw one bar between 41 and 50 then a second one between 51 and 60 on our scale, and leave a gap between the rectangles! Instead, we would redefine our class boundaries in such a way as to eliminate the gap between 50 and 51 (i.e. make the classes continuous). We could change 41 to 50, 51 to 60, etc. into $40 < x \leq 50$, $50 < x \leq 60$, etc. (this keeps all the classes essentially the same, but now makes them continuous). **This is one of the reasons why we are much better off using continuous class limits when making a frequency table.** Otherwise, we are going to end up having to alter our frequency table anyway when it comes time to make a histogram.

<u>Always</u> use some type of <u>continuous</u> class boundaries (where one endpoint is included but the other is not) on a frequency table to set yourself up for making a histogram. If you have already been given a frequency table that does not have continuous class boundaries, and have been told to make a histogram out of it, then first change the frequency table to have continuous class limits before making your histogram.

Solution to Question 4(c)



A scale does not have to start at 0; just start your scale near the minimum value. I have used the "relative frequency" as the vertical scale as this can be more easily compared to other data. It would have also been correct in this case to use simply the "frequency" for the vertical scale. In that case, I would have used a scale of "5", "10", "15", and "20" and labelled it "Number of Students".

Note also that I have labelled the horizontal and vertical axes ("Test Scores" and "Percentage of Students" respectively) and that I have given the histogram a title "Test Scores of Students"). Make sure you do likewise on any histogram you are asked to make.

Finally, if you are ever asked to use *JMP*™ software to generate a histogram, do not let the results confuse you. For one thing, *JMP*™ makes histograms sideways (the bars stick out to the right instead of rising up). Who cares! Let *JMP*™ do what it wants, and be thankful you don't have to make one yourself. (If you want the histogram to have the proper orientation, click the red triangle and select "Horizontal Layout" in "Display Options".) Also, don't worry about labelling the axes or titling the graph or anything. Unless you are specifically told to do such things, leave the graph just the way it automatically appears in the *JMP*™ printout.

4. (d) Construct a stemplot.

To make a stemplot (also called a stem and leaf plot), arrange the data from smallest to largest, then break each piece of data into its stem and its leaf. The leaf is always a single digit (the last digit); the rest of the digits comprise the stem. For example, if we had the data value "123", then the "3" would be the leaf and the "12" would be the stem, like so: $\underbrace{12}_{stem}\ \underbrace{3}_{leaf}$. This also means we need to have 2-digit numbers at least (so that we have the one digit we need for the leaf and at least one digit for the stem). For example, our minimum data value in this problem is 3. Since it has only one digit, that must be the leaf. To have a stem as well, we simply put an understood 0 in front of the 3 (write the number "03", which does not change its meaning in any way). That way we see that 0 is the stem and 3 is the leaf, like so: $\underbrace{0}_{stem}\ \underbrace{3}_{leaf}$.

Here is all the data organized in ascending order and with the leaves underlined:

03 44 44 45 47 50 51 52 52 57 57 57 59 61 61 63 64 65 66 66 66
66 70 72 73 75 75 75 77 77 77 77 77 78 79 79 79 80 80 81 83 84
84 84 86 87 88 92 92 100.

We then present this data in two columns with a dividing line between them. The first column is labelled "Stem" and starts with the stem of your minimum value ("0" from the minimum of "03" in this case) and counts to the stem of your maximum value ("10" from the maximum of "100" in this case). The second column is labelled "Leaf" and shows all the leaves for each stem. Make sure your leaves are nicely lined up in columns, so that it is visually obvious which stems have lots of leaves and which do not. We then simply line all the leaves up for a given stem. For example, we see that four people scored in the 40's: 44, 44, 45, 47, so in the stem column we label "4" and in the leaf column we write all four leaves: 4457 in a string: "4|4457". We know that each leaf has only one digit, so that is not four thousand four hundred and fifty-seven attached to the "4" stem. We know, instead, it is four separate pieces of data with a stem of "4", i.e. 44, 44 again, 45, and 47, the original test scores, of course.

Solution to Question 4(d)

Stem	Leaf
0	3
1	
2	
3	
4	4457
5	01227779
6	113456666
7	023555777778999
8	0013444678
9	22
10	0

We count all stems from the lowest, "0", to the highest, "10", in sequence, **even if a stem has no leaves** (such as "1", "2" and "3" in this stemplot). This enables us to identify outliers at a glance. Here, we can clearly see that 03 (i.e. 3) is an untypically low test score since there is such a gap between it and the next score (44).

Also, note the leaves are lined up neatly in columns so that we can recognize at a glance which stems have lots of data, and which stems do not. For example, we see that a lot of people scored in the 70s while hardly anybody scored in the 90s.

Why make a stemplot?

A stemplot, like a histogram, gives us a visual idea about the shape, centre and spread of the distribution of our data, but, unlike the histogram, we have the added benefit of seeing the actual data itself. For example, we can actually see what test scores people got simply by reading this stemplot (somebody scored 3, two people scored 44, etc.) while that information was lost on our histogram. In the histogram we made back in part (c), we can only see that very few people scored between 0 and 10 (2%), and not many scored between 40 and 50 (8%), etc. **Consequently, where it is feasible, a stemplot is a better presentation of data than a histogram**.

Why make histograms then?

A histogram is useful if there is so much data that it would be too awkward to use a stemplot. Imagine having ten thousand test scores; it would be totally impractical to try to make a stemplot because some stems would probably have thousands of leaves. A histogram would simply condense this down to percentages in each class, giving us something much easier to see and understand. Secondly, histograms are more flexible because we can use anything we want for class limits. We can count by fives, tens, twenties, hundreds, thirteens if we want, whatever. On the other hand, stemplots are pretty much stuck to going by tens (although see part (e) coming up). **Histograms are much more versatile than stemplots. Histograms can condense any quantitative data into an easy to understand graph, no matter how much data we have and how narrow or wide the spread in the scores are.**

Imagine if the data you collected spanned from a lowest score of 23 to a highest score of 659. To make a stemplot, our stems would range from 2 to 65! Imagine trying to fit that all on to one page, having to write stems 2, 3, 4, … 63, 64, 65 all the way down your first column. Or, what if you had fifty pieces of data spanning from a lowest score of 235 to a highest score of 247. That means your stemplot would only have two stems, 23 and 24, but, with fifty pieces of data, you are going to end up with a lot of leaves to squeeze into one or both of those stems. What good is a graph going to be that has just two lines on it? In both of these examples, we could make a histogram instead, adapting the scale to suit the data (or, better yet, let *JMP*™ make one for us).

4. (e) Construct a split stemplot.

A split stemplot is constructed in the same way as a regular stemplot. The difference is only in the way the data is presented. In a regular stemplot (as in (d) above), the data goes by tens; the stems 0, 1, 2, 3, etc. are really **0**0, **1**0, **2**0, **3**0, etc. In a split stemplot, we write each stem twice. The stems will be 0, 0, 1, 1, 2, 2, 3, 3, etc. in this problem. What that means is that we are now going by fives. The stems are really counting **0**0, **0**5, **1**0, **1**5, **2**0, **2**5, **3**0, **3**5, etc. Specifically, the first "0" stem is given leaves with the digits 0 to 4 while the second "0" stem is given leaves with the digits 5 to 9; the first "1" stem is given leaves with the digits 0 to 4 while

the second "1" stem is given leaves with the digits 5 to 9; this pattern repeats for all the stems. **Simply count by fives: 0, 5, 10, 15, 20, 25, etc. This reminds you what the *lowest* digit for the leaves can be for that particular stem.**

Solution to Question 4(e)

Stem	Leaf
0	3
0	
1	
1	
2	
2	
3	
3	
4	44
4	57
5	0122
5	7779
6	1134
6	56666
7	023
7	555777778999
8	0013444
8	678
9	22
9	
10	0

The first "7" stem got far fewer leaves than the second "7". This is because the first "7" stands for **7**0, reminding you it gets the leaves from "0" to "4"; the second "7" stands for **7**5, telling you it gets the "5" to "9" leaves.

I did not bother to write the second "10" stem. This is unnecessary since there are no test scores of 105 or higher. You could include that stem, if you wish (with a blank space in the leaf column, of course).

Why make a split stemplot?

A split stemplot is used if we believe there would not be enough stems to present the data well with a regular stemplot. The goal is always to present data in a way that does not mislead the reader. Again, we can use the $\sqrt{n}$ rule to decide approximately how many stems we would like. Since that rule suggests 7 stems would be about right, a regular stemplot would probably have been fine.

Note that the *JMP*™ software makes stemplots as well (using the "Stem and Leaf" option), but they tend to look pretty different than what you would make by hand. First of all,

JMP™ makes them upside down (starting with the maximum stem and counting down to the minimum; I swear the programmer for *JMP*™ must be some sort of anarchist). Secondly, *JMP*™ may use all sorts of unusual splits if it thinks that is necessary to present a better picture. You might seem a stem repeated 5 times (1, 1, 1, 1, 1, 2, 2, 2, 2, 2, etc.). That would mean *JMP*™ is going by 2s (**1**0, **1**2, **1**4, **1**6, **1**8, **2**0, etc.). Again, who cares!

A regular stemplot writes each stem once.
A split stemplot writes each stem twice.

What if the data for my stemplot is awkward?

In class, you may see all kinds of data and be asked to make a stemplot. Just remember the last digit in each data value is the leaf. By default, we assume the data presented in a stemplot are whole numbers, like we have just seen, but stemplots can be used to display all kinds of numbers. For example, perhaps you recorded data to one decimal place like "12.3". We would then have a stem of "12" and a leaf of "3". You would make your stemplot in the usual way, but include a note with your stemplot showing people how to read it properly (12|3 would look like 123, unless you tell the reader it says 12.3).

All data must be the same kind of number in order to make a stemplot (all whole numbers, all numbers with one decimal place, all numbers with two decimal places, etc.). Of course, any researcher who knows what they are doing, will have measured all their data with equal accuracy anyway, sot this is of no concern. However, if you are ever looking at data where some have decimals and others do not, for example, you would simply alter the data to make it all the same style. Go with the majority. If most are whole numbers, then round the few decimals off into whole numbers as well. If most of the numbers have one decimal place, but you have a whole number like 62 as well, simply add a decimal place (62.0) to make it match up with the rest.

Here are some examples of stemplots you might see, the first stemplot shows data ranging from a minimum value of 30 to a maximum value of 56. Since that is the way a person would read the stemplot anyway, there is no need for an explanation. The second stemplot appears to display data ranging from 51 to 77, but the last digit is really a decimal value, so I

include a note explaining to read the data as 5.1 to 7.7. The third stemplot appears to display numbers ranging from 0 to 24, but they are really from 0 to 2400s, so I include a note explaining to multiply each value in the stemplot by 100 in order to read them properly. There are other ways I could get these messages across (see my footnotes), the key is just to make sure a reader knows what to do.

Stemplot 1

Stem	Leaf
3	00258
4	011136779
5	22246

Stemplot 2: Data ranges from 5.1 to 7.7[*]

Stem	Leaf
5	12229
6	0122455789
7	005567

Stemplot 3: Multiply each value by 100[†]

Stem	Leaf
0	011446
1	01225889
2	0001112334

[*] A more mathematical person might describe **Stemplot 2** this way: Divide each value by 10. So, what appears to values from 51 to 77 are actually $51 \div 10 = 5.1$ to $77 \div 10 = 7.7$. Somebody who knows scientific notation might tell people to multiply each value by 10^{-1}, which is just a fancy way of saying move the decimal one place to the left. **You don't have to be a mathematician or a scientist! Just make sure you have made it clear to people how to read the numbers on your stemplot if they are unusual.** Giving them an example or two of how to read the numbers on the stemplot, like I did above, is perfectly acceptable.

[†] **Stemplot 3** appears to have values from 00 (i.e. 0) to 24, but we are told to multiply each value by 100, so the values are really from 00×100 to 24×100; which is to say, the values are from 0 to 2400. You could, instead, write a note showing people what the numbers really are, like I do in Stemplot 2, if you don't want to be so mathematical. For example, you could say the numbers are in the hundreds, ranging from 00 *hundred* to 24 *hundred*; or, you could tell people to add a couple of 0s to each number, thus 0000 to 2400. There are many ways to get the message across. On the other hand, a show-off might use scientific notation, and tell people to multiply each value by 10^2, which is a scientist's way of saying move the decimal two places to the right.

TRIMMING DATA

If a person is bound and determined to make a stemplot even when the data is not very cooperative, they might **trim** the data. Let's say we had fifty scores ranging from 123 to 968. If we were to leave that as is, we would end up with stems ranging from 12 (with a leaf of 3) to 96 (with a leaf of 8). We would have to making a column listing stems 12, 13, 14 … 94, 95, 96. That is way too many stems! Good luck trying to fit all of them on one page (no magnifying glass allowed). To get the scores under control, we trim the data. Which is to say, we cut away the last digit (as though we trimmed it with a pair of scissors). 123 has its 3 trimmed away, leaving 12; 968 has its 8 trimmed away, leaving 96. It now looks like our numbers are ranging from 12 to 96, but we must remember the 12 is really one hundred and twenty *something* (the 3 is trimmed away and lost forever, so no one will ever know it was one hundred and twenty-three); similarly, the 96 is really nine hundred and sixty *something* (the 8 is trimmed and lost forever).

Now, our trimmed data can be lined up into a stemplot. Since the data now ranges from 12 to 96, we only need stems from 1 to 9. Because the data is really 120 to 960 (again, we will never know what the true last digit was once it has been trimmed), I must tell readers how to convert the numbers on the stemplot. Thus, the stemplot would look something like this (I put * to represent the other 48 leaves we were not given and I just placed them where I felt like in the plot):

Stemplot 4: Multiply each value by 10

Stem	Leaf
1	2*****
2	**********
3	***
4	********
5	******
6	*****
7	****
8	*****
9	*6

Don't confuse trimming data with rounding it off. In the above example, we could round the numbers off to the nearest ten, if we wished. Recall, when rounding off, you round up when the digit is 5 or higher; otherwise, you leave it alone. Then 123 would round off to 120 while 968 would round off to 970. We could then trim the 0s off to have 12 to 97. Some people might consider this better since 968 is almost 970 anyway, while simply trimming 968 makes it become 960 instead. That's not our concern. **If you are told to trim data, do <u>not</u> round it off first.** After all, by trimming or rounding, we are losing the exact values anyway, and nobody cares if the actual number was closer to 960 or 970.

If our quantitative data would require too many stems to make a stemplot viable, we can trim the data (cut away the last digit) and make a stemplot with those numbers instead. If, after we trimmed the data, we still have too many stems, we can trim another digit away, and so on. Be sure to include instructions how to read the data on the stemplot (such as "multiply the numbers by 10").

4. (f) Discuss the shape of the distribution. Are there any outliers?

When describing the shape of a distribution, first, check for outliers. **An outlier is any piece of data that is clearly outside the overall pattern (much smaller than the rest, or much larger).** Then, *ignoring the outliers*, describe the shape of the rest of the distribution by looking at any graph you have drawn. In our case, we have a histogram, stemplot and split stemplot to look at. Usually, you will have just one of these to examine. In general, we want to know how many **peaks** there are, and whether the distribution is **symmetrical** or **skewed**.

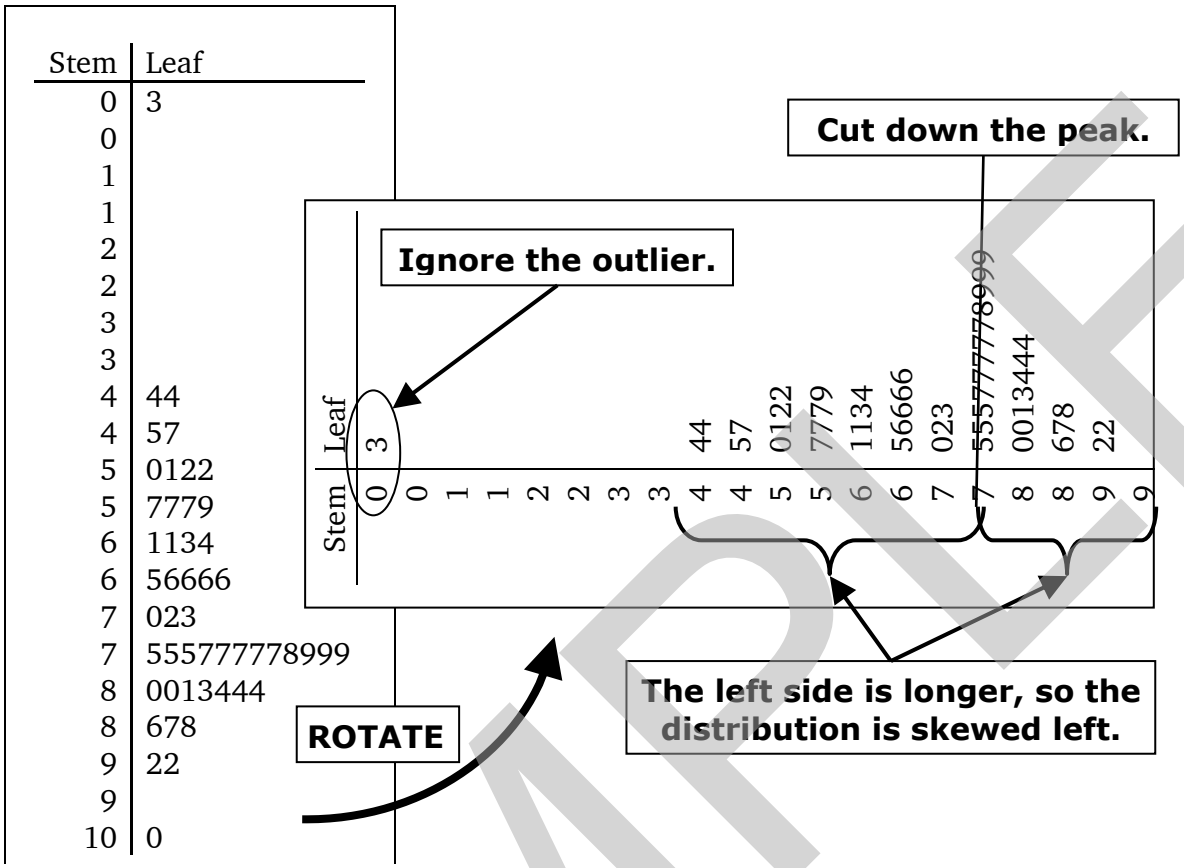
How do we determine the direction of skew?

When looking at a histogram or stemplot, visualize the graph cut down the middle of the peak.[*] If it seems roughly equal in width and shape on both sides of the peak, we will call the distribution symmetric. If one side of the peak looks clearly wider than the other side, we will call the distribution skewed. **If the left side is longer, it is left-skewed. If the right side is longer, it is right-skewed.** Remember, **ignore outliers while determining if there is skewness.** Skewness is determined by the overall pattern of the data, not one or two unusual values.

> **The direction of skew assumes the horizontal axis shows the range of the variable's values. A stemplot runs the variable vertically down the "stem" column. We have to turn our stemplots sideways in order to identify the direction of skew. Be sure you rotate a graph in such a way that the horizontal scale is moving from small numbers on the left to large numbers on the right. Otherwise, you risk confusing the direction of skew.**

Look at the histogram we drew on page 26, the stemplot we constructed on page 28, and the split stemplot we constructed on page 30. I think the split stemplot actually tells us the shape best. Rotating it counterclockwise so that the stems move horizontally from 0 on the left to 10 on the right, we clearly see the stemplot has one peak (the "75" stem), and, even ignoring the outlier at 3, the left side looks much wider than the right so, we would say the distribution is single-peaked and left-skewed with one outlier on the left (3). The regular stemplot and histogram both tell us the same story, but the skewness is less obvious.

[*] If a graph has two peaks of approximately equal size, cut down the middle between the two peaks to check for symmetry or skewness. If one of the two peaks is clearly higher than the other, cut down the middle of the higher peak. Use similar approaches for graphs that have three or more peaks (which would be pretty rare).



Solution to Question 4(f)

The distribution is single-peaked (peaking in the mid-70s) and skewed to the left. The test score of 3 is an outlier.

This is exactly why statisticians have all sorts of graphs. Each problem is a matter of determining what picture and what scale most fairly represents the data. All three of the graphs we drew do a fair job of presenting the information, but the split stemplot certainly helped us see the skewness that was perhaps less obvious in the other two graphs (especially the histogram).

It is unlikely we will have to decide what graph to draw ourselves. They will tell us, "make a stemplot", for example, so we do. Then we will look at that graph to decide the shape of the distribution (and have every right to expect the graph they had as draw will make the shape pretty clear). If on an assignment or exam, you are having trouble deciding on the shape with a regular stemplot, consider quickly making a split stemplot to see if that helps crystallize your thoughts.

4. (g) Find the median test score.

The median is a measure of the centre of a distribution. First, we must arrange the data from smallest to largest (which we already did when we constructed the stemplot). The median is located in such a way that 50% of the data is below the median, and 50% is above the median. We use **the $\frac{n+1}{2}$ rule** to find the *location* of the median.

In our problem, $n = 50$, so $\frac{n+1}{2} = \frac{51}{2} = \mathbf{25.5}$. This tells us the median is halfway between the 25th and 26th *ordered* data value. Start at the minimum value and count to the 25.5 position (I have shown the count below each ordered value in my list):

3	44	44	45	47	50	51	52	52	57	57	57	59	61
1	2	3	4	5	6	7	8	9	10	11	12	13	14

61	63	64	65	66	66	66	66	70	72	73	75	75
15	16	17	18	19	20	21	22	23	24	25	26	27

75	77	etc.
28	29	

The median is the average of 73 and 75.

The 25.5 position takes us between the "73" and "75", so we average these two values to get the median.

$$\frac{73+75}{2} = 74$$

Solution to Question 4(g)

The median test score is 74.

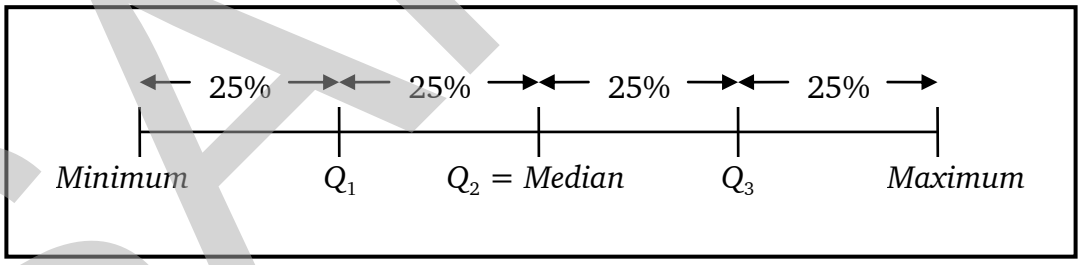
If we had $n = 51$ pieces of data instead (let's assume there was a 51st person who scored 100 on the test), then $\frac{n+1}{2} = \frac{52}{2} = \mathbf{26}$. That would tell us the 26th ordered value is the median. Looking at my count above, I would conclude that 75 is the median in that case (the 26th number). It does not matter the 27th and 28th people also scored 75, I am not even looking at those data values. The 26th number is 75, so the median is 75, end of story (if $n = 51$ of course).

A median is very informative. Here, knowing the median test score is 74, I immediately know half the students scored less than 74 and half scored more. A median is guaranteed to always tell me the halfway point. Make sure you have arranged the data from smallest to largest before finding the median!

If you scored 75 on the test, you know you are in the top half of the class since you are above the median. You also know you are barely in the top half though, so if your prof is marking on the curve, 75 might only be a C or C+.

4. (h) Find the first and third quartiles.

The quartiles of a distribution, as the name implies, break the *ordered* data into 4 equal pieces. Each piece is 25% of the data (see the diagram below). **The first quartile, Q_1, is located such that 25% of the data is below it (75% is above it); the second quartile, Q_2, is actually the median (50% of the data above and below it); the third quartile, Q_3, is located such that 75% of the data is below it (25% is above it).**



TO FIND THE FIRST AND THIRD QUARTILES:

Step 1: Arrange the data from smallest to largest, and use $\frac{n+1}{2}$ to locate the median. **Visualize a cleaver slicing through the data at the median, cutting it into two equal pieces. If the cleaver would actually strike a piece of data, pretend that piece is smashed to smithereens.** How much data is in each half?

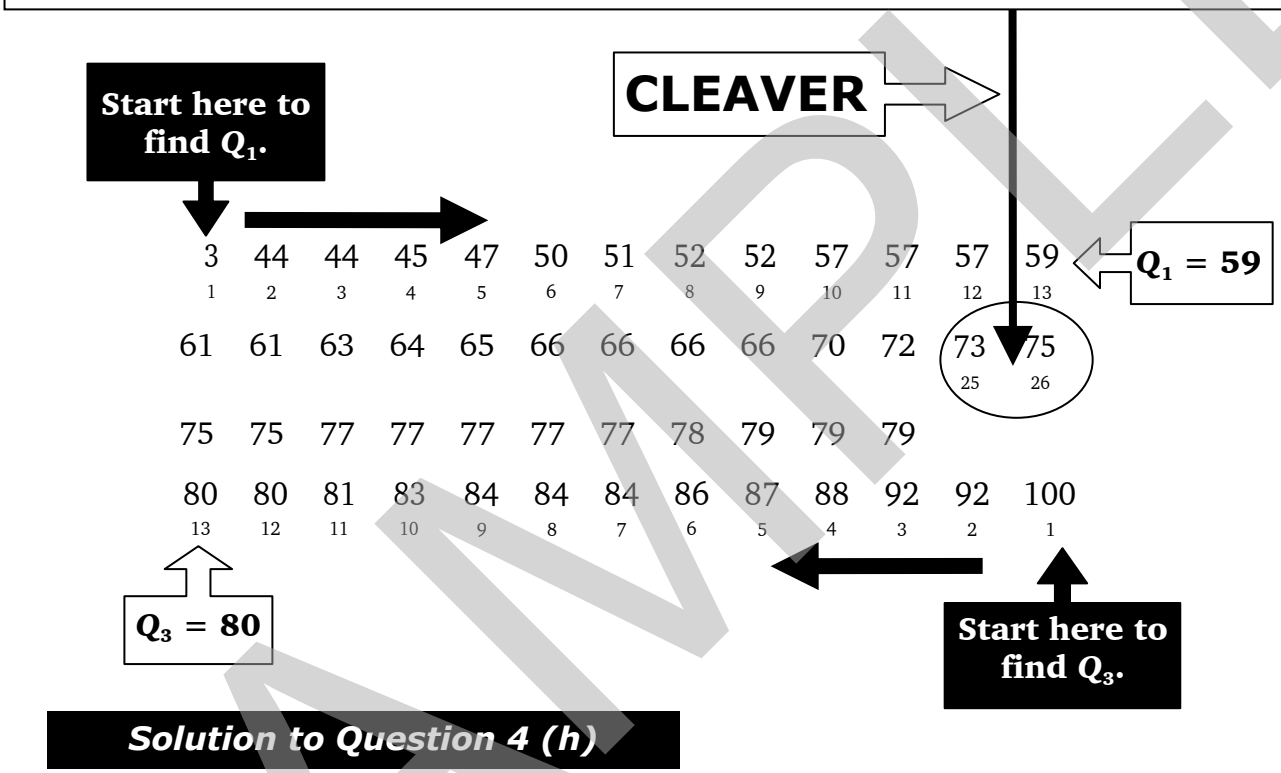
Step 2: Now cut each half of the data in half again by letting "n" equal how much data is in each half. Which is to say, let n = the amount of data below the cleaver (which is also the amount of data above the cleaver). **Remember, if the cleaver hit a piece of data, it is in neither half (it is smashed to smithereens).**

Step 3: Use $\frac{n+1}{2}$ for this new "n" value to find the location of Q_1 and Q_3 in their respective halves (we are finding the median of each half).

Step 4: You can now count to that value in each half to find the first and third quartiles. It is usually easiest to start from the *Minimum* and count forward "▶" to find Q_1, and start from the *Maximum* and count backward "◀" to find Q_3. **Which is to say, count from the far left to find Q_1 and count from the far right to find Q_3.**

Recall, in part (g) we found the median is at the 25.5 position. So, we take our cleaver and slice between the 25th and 26th number. That means the cleaver didn't actually touch a piece of data (no one got smashed to smithereens). Thus, there are 25 pieces of data below the median (below the cleaver), as well as 25 pieces of data above the median. Use $n = 25$ to find the first and third quartiles.

$$\frac{n+1}{2} = \frac{26}{2} = 13 \rightarrow \text{The 13th value in each half is } Q_1 \text{ and } Q_3.$$



Solution to Question 4 (h)

The first quartile = Q_1 = 59 and the third quartile = Q_3 = 80.

The quartiles help us get a better picture of how <u>spread out</u> the data is. For example, we just found in part (g) that the median test score is 74. We know half scored below 74 and half above, but that does not tell us *how close* to 74 people typically scored. There are so many possible mark distributions that could have had a median of 74. Perhaps all fifty students scored 74! That would certainly make the median 74. Perhaps twenty-four students scored 0, one student got 73, one student got 75, and twenty-four students scored 100. That would still mean we would have averaged 73 and 75 (the 25th and 26th numbers) and got a median of 74. Of course, it would seem more likely that there was a wide variety of

marks from very low to very high which ended up centering around 74. But that is just the point! We should not have to assume what the data was like. **A statistical summary of the data should give people a good idea of the mark distribution without having to resort to looking through all the data themselves.**

Now, with the median and the quartiles, we have a pretty good idea about the mark distribution. Only 25% of the class scored less than 59 (the first quartile) while 25% of the class scored better than 80 (the third quartile). Half of the class (the middle 50%) scored between 59 and 80. If I were a teacher for this class, where a fail was any mark below 50, 60 was a C, 70 a B, 80 an A, and 90 were an A+, I think I would be pretty happy with this distribution. It looks like 25% of the class are getting an A or better; half the class scored above 74 (the median) so there will be a lot of B's as well. Since the first quartile is 59, pretty much 75% of the class are getting no worse than a C.

4. (i) Find the Range and Interquartile Range.

Range and Interquartile Range are measures of the spread of a distribution. The Range is simply the difference between the Minimum data value and Maximum data value. The Interquartile Range (*IQR*) is the difference between the first and third quartiles (giving us an idea of how spread out the middle 50% of the data is). **The *IQR* is a more reliable measure of spread than the Range since outliers can affect the range.** For this problem:

Solution to Question 4(i)

$$\textbf{\textit{Range}} = \textbf{\textit{Max}} - \textbf{\textit{Min}} = 100 - 3 = 97$$
$$\textbf{\textit{IQR}} = Q_3 - Q_1 = 80 - 59 = 21$$

In this problem we see there is quite a large spread between all the marks (the range is 97), but the middle 50% of the students have a considerably smaller spread (the *IQR* is only 21). The range gives the appearance that there was quite a large variety of marks (after all the maximum range would be 100 since the lowest possible mark is 0 and the highest possible is 100), but the marks were not as spread out as that suggests. The outlier value of 3 has made

the range quite deceptive. The second lowest test score was 44. If that had been the minimum value, the range would have been 100 – 44 = 56, considerably less.

We can certainly see how the presence of an outlier can cause the range to give a misleading sense of the spread of a distribution. The Interquartile range is immune from this effect since it deals only with the spread of the middle 50% of the distribution.

Some students get confused when there is an outlier and wonder if it should, perhaps, be ignored in calculations. For example, should the Range actually be 56 in this problem (the amount we would have got if we had discarded the outlier value of 3)? Absolutely not! **Never discard outliers unless specifically told to do so!** Range is *Max – Min*, end of story! If the *Max* or *Min* are outliers, so be it. Here, the Range is 97, period.

4. (j) State the five-number summary.

The 5-number summary for any distribution is:

Minimum, First Quartile, Median, Third Quartile, Maximum

Solution to Question 4(j)

The five-number summary is 3, 59, 74, 80, 100.

The 5-number summary gives a nice snapshot of the centre and spread of a distribution. We of course, get all the benefits of knowing the median and quartiles, but also can see the minimum and maximum value. In this problem, we have honed fifty pieces of data down to five key numbers.

4. (k) Justify the outliers (if any) mathematically.

Generally, we will merely identify outliers visually from stemplots and other graphs as we have done already. If we are ever asked to justify our observations, we can use **the 1.5 × *IQR* rule**. As the name implies, we first compute the *IQR* (interquartile range) then multiply that by 1.5. Any piece of data that is more than $1.5 \times IQR$ below Q_1 (the first quartile) or above Q_3 (the third quartile) is considered an outlier.

In part (i) above we found the interquartile range was 21, so $1.5 \times IQR = 1.5 \times 21 =$

31.5. Now, subtract 31.5 from Q_1 and add 31.5 to Q_3. In part (h) we found $Q_1 = 59$ and $Q_3 = 80$, so $59 - 31.5 =$ **27.5** and $80 + 31.5 =$ **111.5**.

Solution to Question 4(k)

Anything below 27.5 or above 111.5 is an outlier. We see that "3" is the only test score below 27.5, and not even "100" (the maximum) is above 111.5. Consequently, our earlier observation that 3 is the only outlier has been justified mathematically.

Don't think you have to do this. If they ask you on an exam or assignment, "Are there any outliers?", you are expected to simply look at whatever picture of the data you have available and see if there are any values clearly sticking out from the overall pattern. Only if they ask you to justify the existence of outliers are you obliged to use the $1.5 \times IQR$ rule.

I generally use what I call the "**two-gap rule**". If there are at least two gaps in a diagram between a piece of data and the rest of the data cluster, then that is enough to convince me there is an outlier. For example, if we look at our split stemplot in part (e) above, "3" is so clearly an outlier. There are seven gaps (seven empty stems) before we finally reach the cluster of data that begins with 44. The "100" I would not consider an outlier since there is only one gap (one empty stem) between it and the cluster.

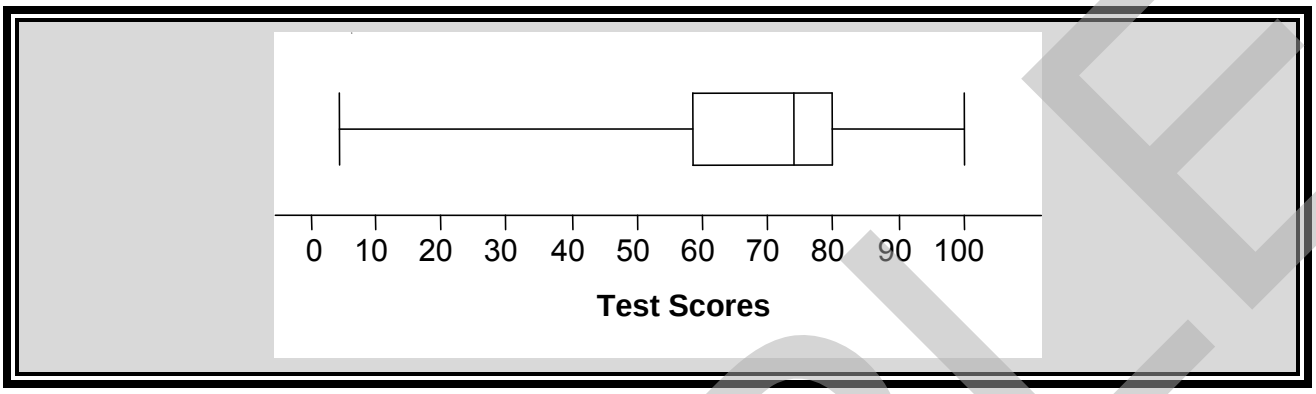
If you are unsure on an exam or assignment, and if there is time, you could use the $1.5 \times IQR$ rule to help you decide if there are any outliers. But try to avoid doing this as it is time-consuming (especially if you haven't even worked out the quartiles in the problem). Only if I have already found or been given the quartiles would I consider using this rule.

4. (l) Draw a boxplot.

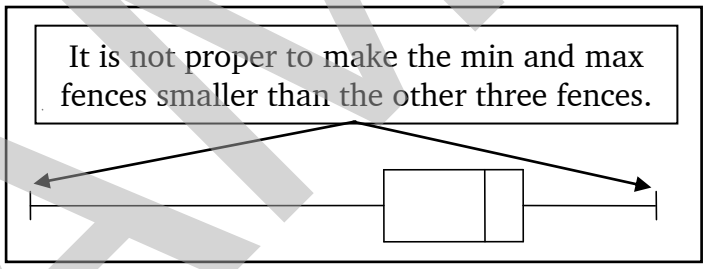
A standard boxplot (which *JMP*™ refers to as a quantile boxplot) is merely a picture of the 5-number summary. After selecting a scale, we draw 5 marker lines (or "fences") at the positions of the 5 numbers in our summary. We then draw a box from Q_1 to Q_3 (the median line would then be somewhere inside the box). **The box in the boxplot shows you the interquartile range, the spread of the middle 50% of the data.** We also draw two "whiskers" out from the box to the lines marking the minimum and maximum value. **I have**

chosen to draw the boxplot horizontally, but boxplots can also be drawn vertically.

Solution to Question 4(l)



I have drawn vertical fences at 3 (the Minimum), 59 (the first quartile), 74 (the Median), 80 (the third quartile), and 100 (the Maximum); a box is then drawn to connect the Q_1 and Q_3 fences together, giving us a visual sense of the spread of the interquartile range. The whiskers have then been drawn out to the fences at the minimum and maximum. **Make sure all five fences are the same size.** Don't do this:

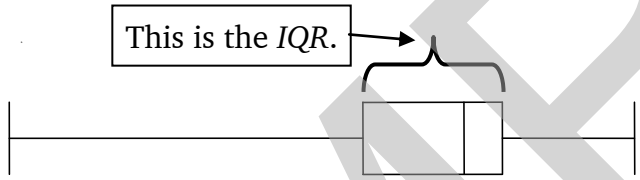


If using a boxplot to determine skewness (if any), look at the box first, not the whiskers. Using the median line as our measure of the centre of the distribution, is the distance to the first or third quartile at the edge of the box approximately the same (suggesting the distribution is symmetric), or is one obviously longer (suggesting a skewed distribution)? We hope the whisker lengths back up our observations in the box but, if not, **we use the box not the whiskers to identify skewness**. The minimum or maximum value might be an outlier causing the length of a whisker to be disproportionately longer than it would have been if the outlier were ignored. Consequently, whiskers can be a misleading depiction of skewness.

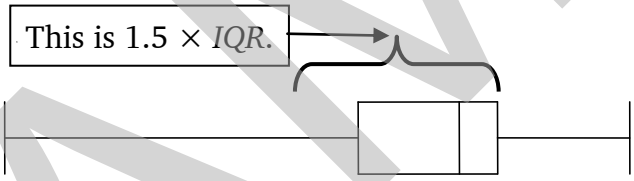
When looking at a boxplot, you can visually use the 1.5 × *IQR* rule to see if the whiskers are being stretched too long by outliers. Stretch your fingers the length of the box (the box's length is the *IQR*). Now, stretch them out half as much again (to estimate 1.5 × *IQR*). If a whisker is any longer than that, you know there are outliers in the data, and the whiskers cannot be trusted to help you identify skewness.

Let's use the boxplot we just made as an example:

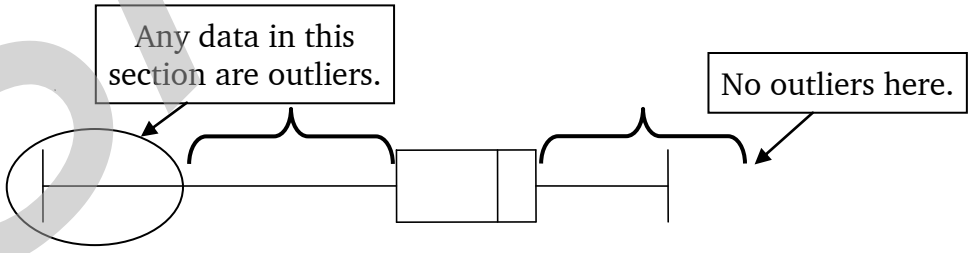
Step 1: Stretch your fingers out the length of the box.



Step 2: Stretch your fingers out half as much again.



Step 3: See how that compares to the whiskers.



The minimum is certainly an outlier. The problem with a boxplot is we have no idea whether any other values besides the minimum are also outliers.

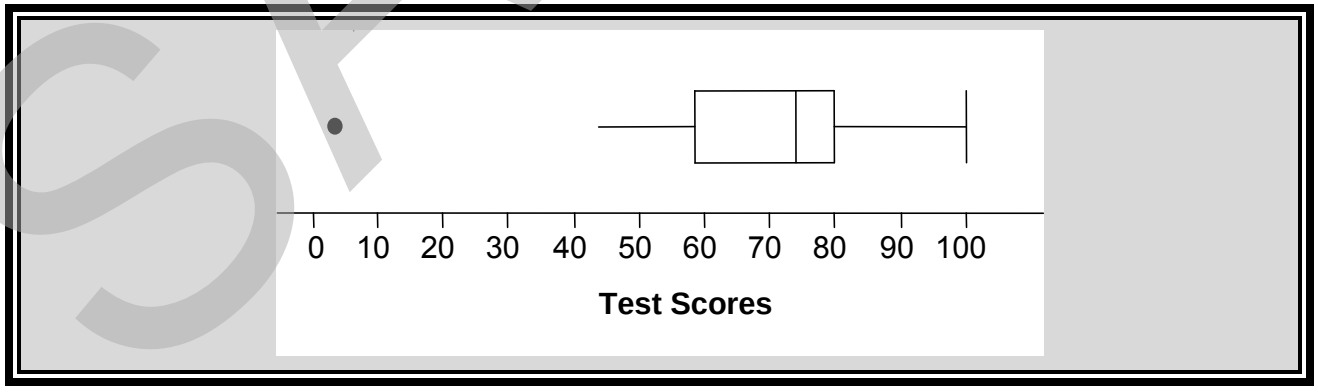
4. (m) Draw a modified (outlier) boxplot.

In our boxplot, we certainly see the box is left-skewed (the part of the box to the left of the median line is much longer than the part to the right of the median). Our whiskers also suggest a left-skew since the left whisker is much longer than the right. This may be misleading though, because we know there is a left outlier (3).

This is where it can be helpful to draw a modified boxplot (which *JMP*™ calls an outlier boxplot). **A <u>modified</u> boxplot or <u>outlier</u> boxplot is precisely the same as a regular boxplot, except any outliers are plotted as dots on the diagram** (if you believe a set of data has four outliers, you will plot 4 dots on your chart to denote them). **We then draw the whisker only as far as the smallest and largest values that are not considered outliers.**

Often, when asked to make a modified boxplot, we are expected to use the $1.5 \times IQR$ rule to establish the outliers. In part (k) above, we established any score below 27.5 or above 111.5 is an outlier. **Do not make the mistake of drawing the whiskers out to 27.5 and 111.5!** Whiskers extend out only as far as the actual data values themselves. We want to draw our left whisker all the way to 3, the minimum, but it was an outlier, so we plot a dot at 3 to signify it is an outlier. Instead, draw the whisker out to 44, the next smallest data value. Additionally, we do not draw a fence at "44" since that may confuse a viewer into thinking 44 is the minimum. Just leave the whisker dangling at 44 with no fence capping it.

Solution to Question 4(m)



Note, the whiskers now look pretty symmetrical, but we would still consider this distribution left-skewed since the box itself is skewed left (the distance between Q_1 and the median is much longer than the distance between Q_3 and the median).

4. (n) Find the mode of the distribution.

This is a rarely used measure of the centre of a distribution. Quite simply, **the mode is the most frequently occurring data value.** In our problem, **the mode (or modal value) is 77**, because that occurred 5 times (more than any other value). We can say that more students scored 77 on the test than any other score.

Solution to Question 4(n)

The mode is 77.

If we were given a histogram to look at, but not the actual data, we can at least identify the **modal class**. **The class with the highest peak obviously has the highest frequency, and so is the modal class.** In our problem, looking at the histogram we made back in part (c), we could say the modal class is "70 – 80". More students scored in the 70's than any other mark range.

4. (o) Find the mean of the distribution.

The mean (or average) of a distribution is another measure of the centre of a distribution. Unlike the median, we do not have to put the data in order. **The mean of a sample is denoted $\bar{x}$** (pronounced "*x*-bar", memorize this symbol!) and is found by adding all the data together and dividing by *n*, the sample size.

We denote addition in statistics by the "Summation" symbol "Σ". (This symbol is called "sigma"; it is the Greek capital "S" for Summation.) For example, $\sum_{i=1}^{n} x_i$ says sum up all the x_i values, starting when $i = 1$ and ending when $i = n$ (i.e. we sub in 1, 2, 3, … *n* in place of the *i* subscript). Which is to say: $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$ where x_1, x_2, x_3, etc. stands for the 1st

piece of data, 2nd piece of data, 3rd piece of data, etc. We generally simplify this notation by simply writing "$\sum x$" (pronounced "sigma x") to denote the sum of all the x-values (the sum of all the data values).

All this boils down to saying: $\boxed{\bar{x} = \frac{\sum x}{n}}$ (**memorize** this formula)

That, of course, just tells us, in order to find the mean of our data, $\bar{x}$, we simply add up all the data values in our sample, $\sum x$, then divide by the sample size, n.

Know how to compute a mean by hand, but never actually bother doing it that way. Your calculator will be able to compute the mean for you by simply inputting the data. BE SURE YOU KNOW HOW TO USE *STAT MODE* ON YOUR CALCULATOR! Check Appendix A at the end of this book for the steps to use for your make and model of calculator. Admittedly, some calculators can't handle 50 pieces of data. If yours lets you down, skip this question. You would never have such a large sample size on an exam question anyway.

For our problem, we have $n = 50$, and we can put our calculator in *stat mode* and simply enter the data in the order it was originally given (the first test score was 75, the second was 88, the third was 47, all the way to the last score of 52). We determine:

$$\bar{x} = \frac{\sum x}{n} = \frac{75 + 88 + 47 + \ldots + 52}{50} = \mathbf{69.14}.^{*}$$

Do not round this number off! **A good rule of thumb is never round numbers off.** If it is clear your calculator is giving you an exact number (i.e. it is not filling the whole screen with digits), then keep the entire number. That is what happened here. The calculator says the answer is exactly 69.14, so I give that as my answer. I do not round it off to 69.1 or 69! **Only if you have an endless decimal value showing on your calculator will it be necessary to round off. My motto is round to 4 decimal places, no less.**

[*] Do not actually add these numbers up! Enter the data using the *stat mode* on your calculator and let it do the computation for you. Keep the data saved in your calculator ready for the standard deviation in the next part.

Solution to Question 4(o)

The mean of the sample is $\bar{x}$ = 69.14.

Essentially, a mean is telling us, if we were to add up all the data, then share it out equally with everyone, everyone would get the mean value. Here, since the mean is 69.14, that is saying the total of all 50 test scores is as if everyone scored 69.14 on the test. If we don't know the original scores, we have no idea how many people scored exactly 69.14, how many scored more, and how many scored less. We can only say, if everyone scored the same, they would have scored 69.14.

The mean of the sample (69.14) is smaller than the median (74). This was caused by the left outlier (3) and the left-skew. **If a distribution is skewed or has outliers, the mean is not a good measure of the centre, because it will be pulled away from the centre in the direction of the skew or outliers.** Here, the mean was pulled to the left of centre by the left-skew. **Medians are resistant to skewness or outliers, and so, are better measures of the centre in these cases.** You always know half the data is on either side of the median.

Here, if I told you the mean test score is 69.14, and you scored 69 on the test, you would probably think about half the people scored higher than you, but that isn't true. We know the median is 74, so a score of 69.14 is certainly not the midway point. Considerably more than half the class scored better than you. The mean is misleading. In fact, if you check the stemplot we made earlier, 28 students scored better than 69.14, the mean. 28/50 = 0.56. So 56% of the class scored better, and only 44% scored worse. A mark of 69 is pretty solidly in the bottom half of the class. The left skew and the outlier on the left (3) have pulled the mean to the left of the median (pulling the mean into the bottom half of the test scores).

In general, if a distribution is symmetric, the mean and median will be the same. In these cases, either could be used to measure the centre, but we prefer to use the mean as it is more useful for other calculations. If a distribution is skewed or has outliers, the mean will be pulled in that direction, away from the median causing the mean to be a poor measure of the centre. In these cases, the median is preferred.

4. (p) Find the standard deviation of the distribution.

The standard deviation (like the range and interquartile range) is another measure of the spread of a distribution. We first compute each data value's *deviation* from the mean "$x_i - \bar{x}$", then square these deviations "$(x_i - \bar{x})^2$, sum them up "$\sum(x_i - \bar{x})^2$", and divide that sum by "$n - 1$". At this point we have found what is called the **variance**. We finally take the square root of the variance to get the standard deviation. **The standard deviation of a sample is denoted s, whereas the variance is denoted s^2 because it is simply the square of the standard deviation.**

We don't really have much use for the variance, it is the standard deviation we really want. Frequently on exams they will mention variance though just to see if you know what it is. Remember the symbols, they say it all:

Standard deviation is s; if you square that value, you get the variance, s^2.

Variance is s^2; if you square root that value ($\sqrt{s^2}$), you get the standard deviation, s.

Memorize these formulas (you may need them on an exam):

$$\textbf{variance} = s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \quad \text{also written } s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

$$\textbf{standard deviation} = s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \quad \text{also written } s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

Never actually use this formula to compute s. Instead, use the *stat mode* on your calculator (see Appendix A) to compute the standard deviation.

Recall, $n = 50$ and $\bar{x} = 69.14$ (found in part (n) above) for this problem, and we do not have to order the data to compute *s*. The work I show below is just to show you what the formula is telling me to do. I wouldn't dream of actually going through all those computations (or is that more of a nightmare?). Thank-you for calculators with *stat mode*! If you already fed your data into your calculator in *stat mode* to get the mean in part (n), you don't have to start all over again. Simply ask the calculator for the standard deviation right away.

Recall, the test score data given at the start of this question was 75, 88, 47, … 52.

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

$$= \sqrt{\frac{\left[(75-69.14)^2 + (88-69.14)^2 + (47-69.14)^2 + \ldots + (52-69.14)^2\right]}{49}}$$

$$= \mathbf{16.77341191\ldots}$$

Round the answer off to 4 decimal places (never less unless they tell you to).

Solution to Question 4(p)

The standard deviation is $s = 16.7734$.

On a hand-in assignment, you will probably be asked to compute a standard deviation by hand, showing all your work. That question will certainly not involve 50 pieces of data. My question 6 below shows you how you would deal with such a question.

Now that we know the standard deviation, we get a sense as to how spread out the distribution is. A small standard deviation suggests a small spread; a large standard deviation is obviously suggesting a large spread. **If all the data had the same value** (if everybody

scored 75 on the test, for example), **then there would be no spread at all, and the standard deviation would be 0.**

We expect a majority (but certainly not all) of the data to be within one standard deviation of the mean. Which is to say, if we compute $\bar{x} \pm s$, we will get a range that includes a majority of the data. Let's round off the mean and standard deviation to 69.1 and 16.8, respectively, to get an idea of what I mean. (I know I said earlier never round numbers off, but shut up! It's my book, and I'll do what I want. If I round to one decimal place, that is using more digits than the original data and should be good enough.) Essentially, we expect a good deal of the test scores are within 16.8 units of 69.1: $\bar{x} \pm s = 69.1 \pm 16$ which gives us the limits of (52.3, 85.9).

Certainly, a great deal of the scores are between 52.3 and 85.9, as a glance at our stemplot would confirm. Essentially, if you know the mean and standard deviation, you get an idea of what a typical value in the distribution would be. Here, I expect a typical student might have scored between 52.3 and 85.9 on the test (i.e. 69.1 give or take 16.8). It is similar to knowing the quartiles of a distribution, but not the same thing.

If we want to get really fancy, we can actually see how much of the scores do lie between 52.3 and 85.9. Looking at the stemplot we made earlier, I count nine values below 52.3 and six values above 85.9 for a total of fifteen values outside of this range. That means thirty-five out of the fifty test scores or 35/50 = 0.7 = 70% of the scores are within one standard deviation of the mean. We, of course, were not asked to do this, but I thought it might be fun (yeah, I'm weird).

However, the **standard deviation is not a good measure of the spread if a distribution is skewed or has outliers**. Firstly, we need to use the mean in order to calculate standard deviation, and we have already pointed out that a mean is not a good measure of centre for skewed distributions. Additionally, skewness or outliers tend to cause a disproportionately large standard deviation, exaggerating the spread. When a distribution is skewed or has outliers, the interquartile range is a more reliable measure of spread, because, like the median, **quartiles are resistant to skewness and outliers**.

In general, for symmetric distributions we prefer to compute the mean and standard deviation as a measure of centre and spread. For distributions that are skewed or have outliers, the five-number summary is best as the median and interquartile range are a more reliable measure of centre and spread because they are resistant to skewness and outliers.

Medians and quartiles are resistant to skewness and outliers; means and standard deviations are not. If we discovered that the "3" in our data above was actually a typographical error, and the real score was actually "43", our mean and standard deviation would change markedly in value. (Try it: the mean changes from 69.14 to 69.94, almost a full digit larger (still lower than the median, though, because of the left skew), and the standard deviation changes from 16.7734 to 14.3533, almost 2.5 units smaller!) Yet, our median and quartiles would stay exactly the same since 43 would still be the minimum value. Since medians and quartiles focus in on the middle of the ordered data, they are completely unaffected by unusually large or small values. Means and standard deviations, however, are very much affected by such values.

SHAPE, CENTRE, AND SPREAD

The ordeal that was question 4 was an exercise in the various things we can do to summarize a quantitative distribution. The three things we should refer to when describing a quantitative distribution are the shape, centre and spread.

SHAPE: **Use graphs (histograms, stemplots and/or boxplots) to visualize the distribution of a quantitative variable. How many peaks does it have? Is it symmetric or skewed? Does it have any outliers?**

CENTRE: **The three measures of centre are mean, median and mode. The mean will be pulled away from the centre in the direction of the skew and/or outliers (if any) while a median is resistant to skewness and outliers. Consequently, a median is a more trustworthy measure of centre when distributions are skewed or have outliers. You always know half the scores are below the median and half are above. When a distribution is symmetric, the mean and median will be the same. The mode tells us what score happens most frequently.**

SPREAD: **If you are using the mean to measure centre, use the standard deviation to measure spread; if you are using the median to measure centre, use the interquartile range and range to measure spread.**

5. (a) Find the first, second and third quartiles for the data set below.

3, 8, 84, 51, 23, 13, 18, 15, 18, 4, 16, 4, 9

See Figure 1 below for an illustration of the work I am doing to find the three quartiles in this problem. Of course, **the second quartile is the median**, so start there.

First and foremost, put the data in ascending order:

3, 4, 4, 8, 9, 13, 15, 16, 18, 18, 23, 51, 84

$\boldsymbol{n = 13} \rightarrow \frac{n+1}{2} = \frac{14}{2} = 7 \rightarrow$ the 7th value is the median or the second quartile, Q_2.

We take our cleaver and slice it through the 7th number. $\rightarrow$ **The median is 15.**

Since our cleaver has smashed the 7th number to smithereens, that leaves 6 numbers below the median and 6 numbers above it. Use $n = 6$ to find the quartiles.

$$n = 6 \rightarrow \frac{n+1}{2} = \frac{7}{2} = 3.5$$

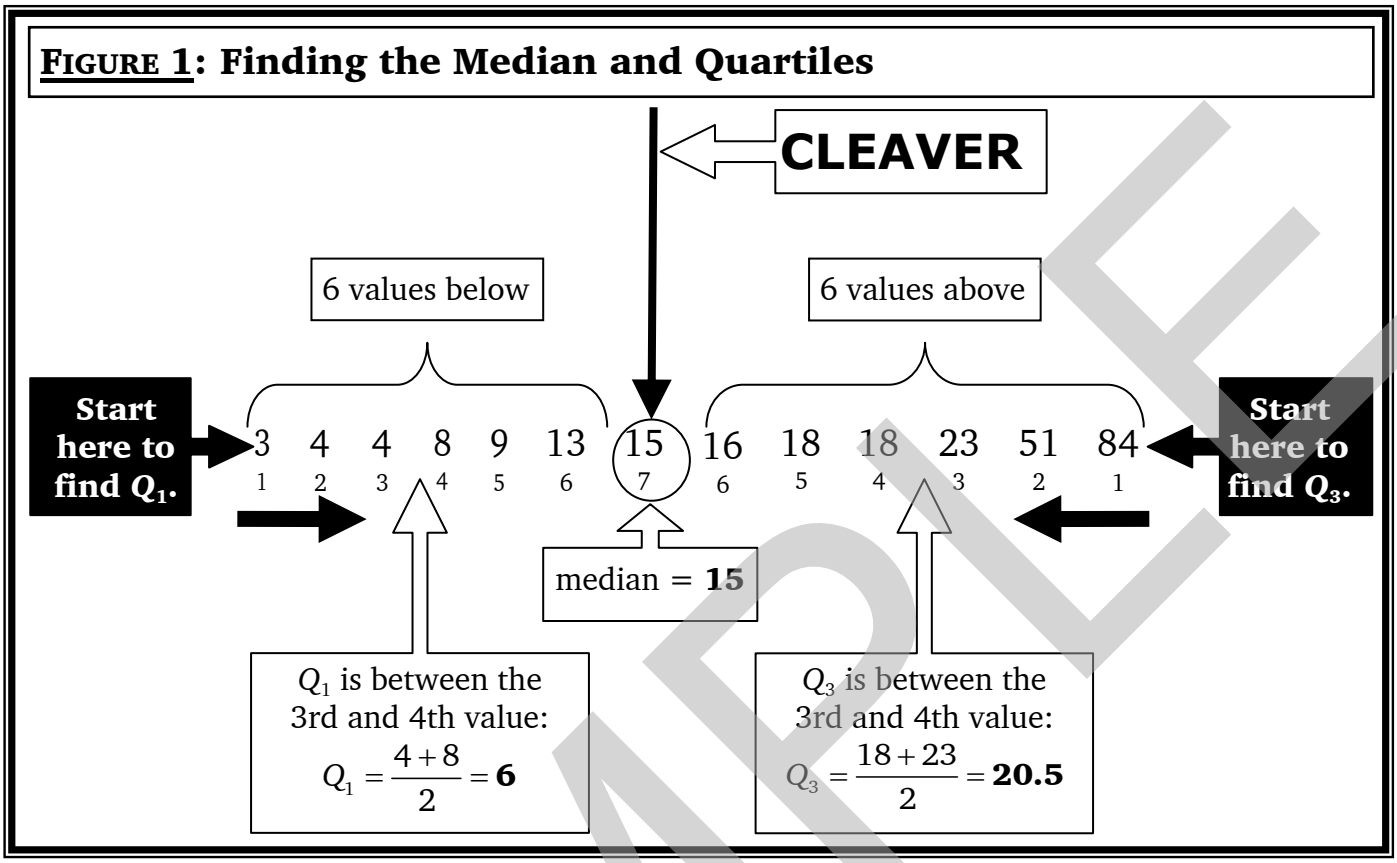
Q_1 and Q_3 lie between the 3rd and 4th number in their respective halves. Averaging the 3rd and 4th number in each half we discover:

$$Q_1 = \frac{4+8}{2} = \mathbf{6} \text{ and } Q_3 = \frac{18+23}{2} = \mathbf{20.5}$$

The first quartile is 6 and the third quartile is 20.5.

Solution to Question 5(a)

The first, second and third quartiles are 6, 15 and 20.5, respectively.



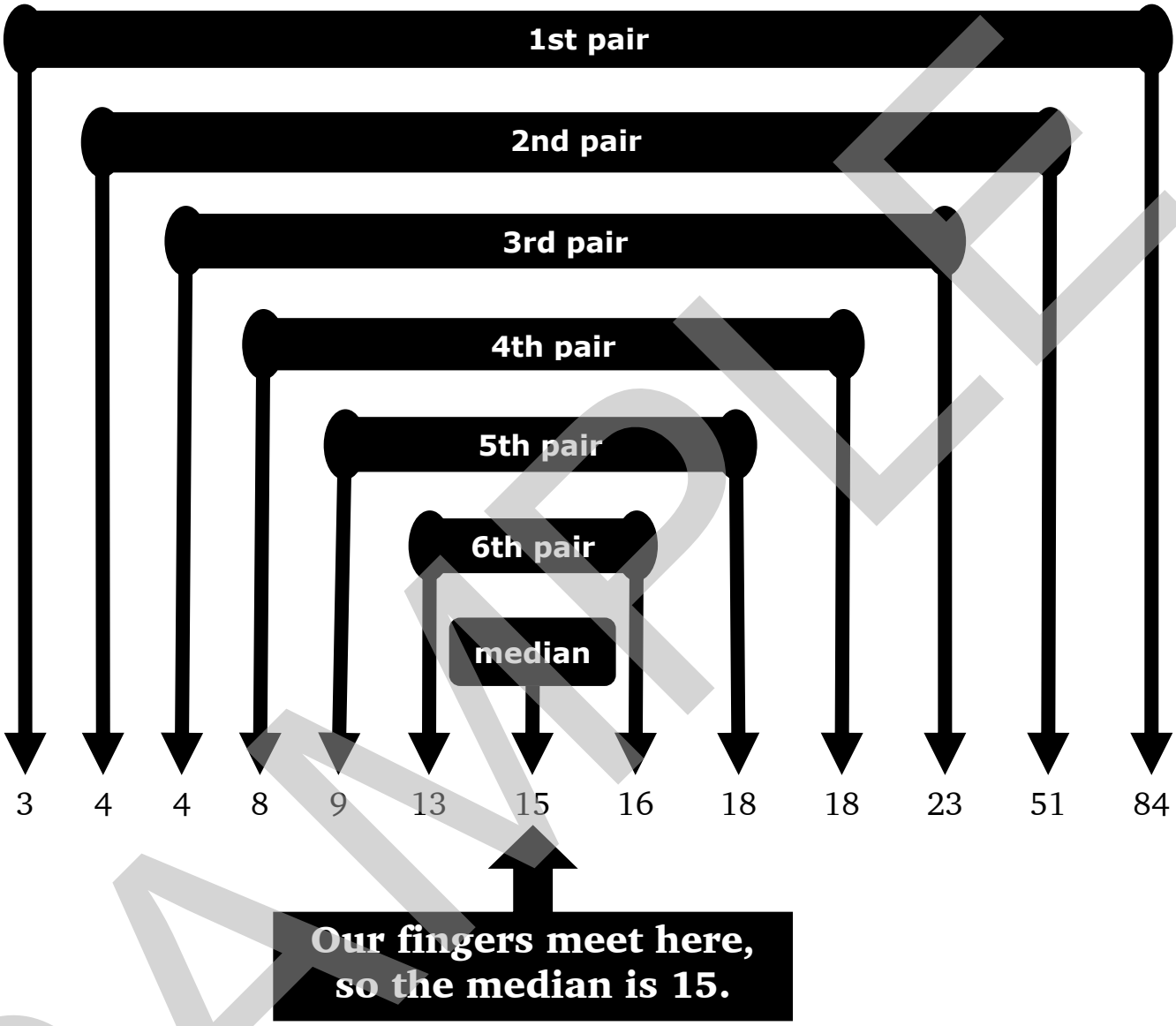
The Finger Method to find the median and quartiles

If you have a reasonably small amount of data (like the 13 values we just saw in the question above), you may find it faster to use your fingers to find the median and/or quartiles. This is a good idea on exam questions since they frequently give you a small amount of data to work with.

To find the median once you have ordered the data from smallest to largest, simply put a finger on each of the first and last number in the list. Now, move your fingers to the second and second last number. Then, move them to the third and third last number, and so on. If you keep pairing the numbers up in this way, working in from both ends, your fingers will meet at the median. Once you have found the median, you can again use a cleaver to slice the data into two halves and use the finger method on each half to count your way to the quartiles, if required. The nice thing about the Finger Method is you don't even have to know the exact amount of data (you don't care what n equals). As long as you have put the data in ascending order, your fingers will guide you to the median (or quartiles).

See Figures 2 and 3 below for an illustration of the Finger Method.

FIGURE 2: Finding the Median using the Finger Method



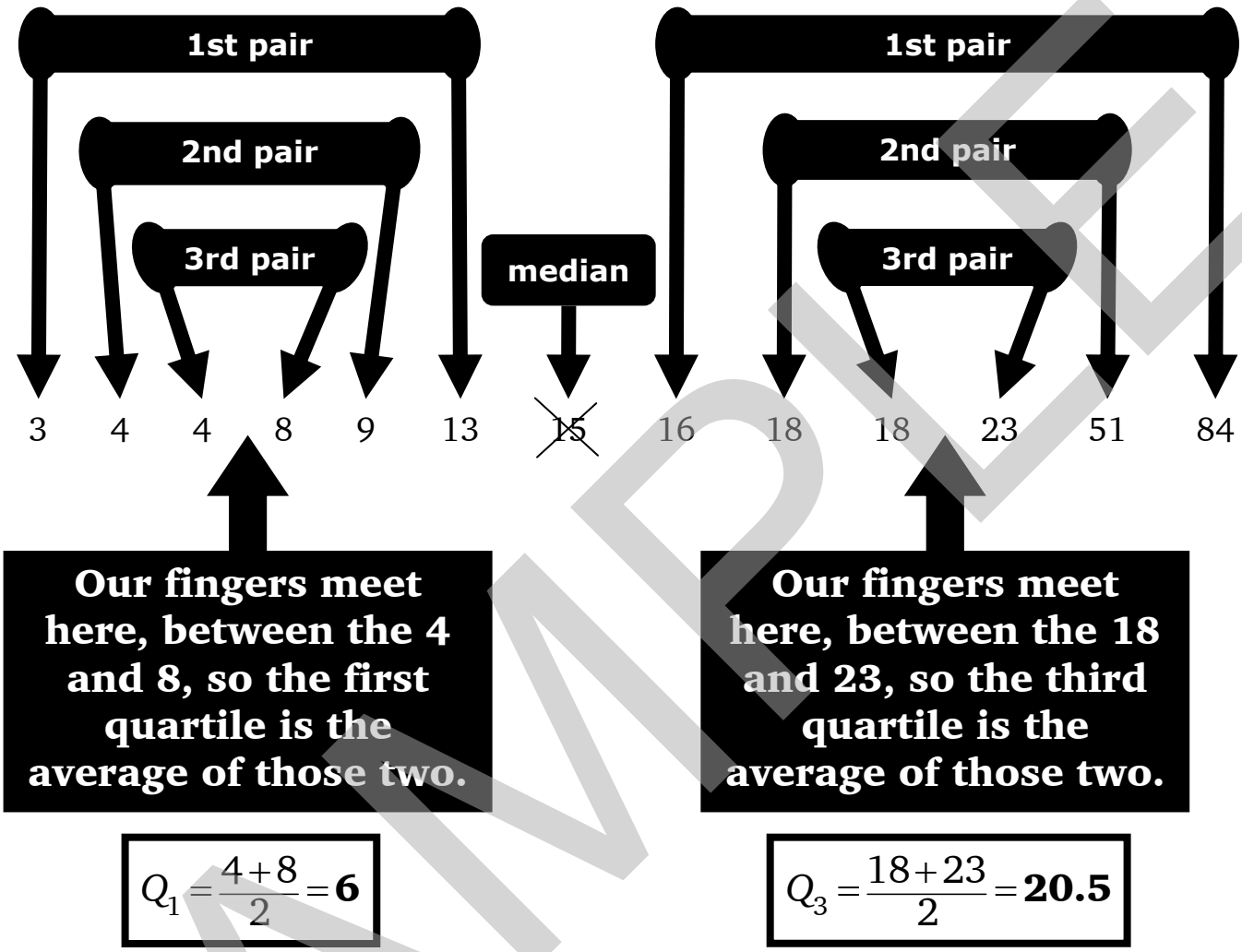
Now that we have found the median, our cleaver smashes that number (15) to smithereens, leaving us to find the quartiles of the remaining halves.

3 4 4 8 9 13 ~~15~~ 16 18 18 23 51 84

Use your fingers to get the first quartile (the median of this lower half of the data set).
See Figure 3 below.

Use your fingers to get the third quartile (the median of this upper half of the data set).
See Figure 3 below.

FIGURE 3: Finding the Quartiles using the Finger Method



The Finger Method confirms:

The first, second and third quartiles are 6, 15 and 20.5, respectively.

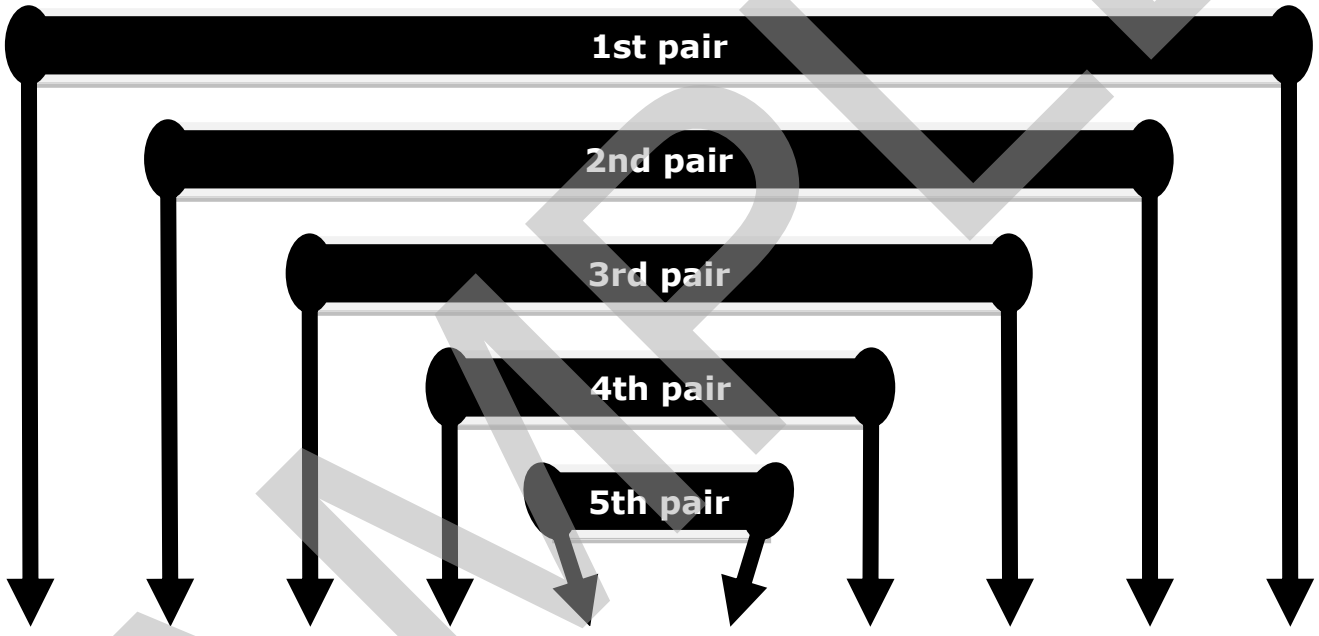
5. (b) Find the first, second and third quartiles for the data set below.

24, 27, 26, 22, 23, 27, 22, 18, 21, 10

Since this is a small data set, I will use the **Finger Method**. Note, I first **arrange the data in ascending order.**

See Figures 4 and 5 below for an illustration of the Finger Method.

FIGURE 4: Finding the Median using the Finger Method



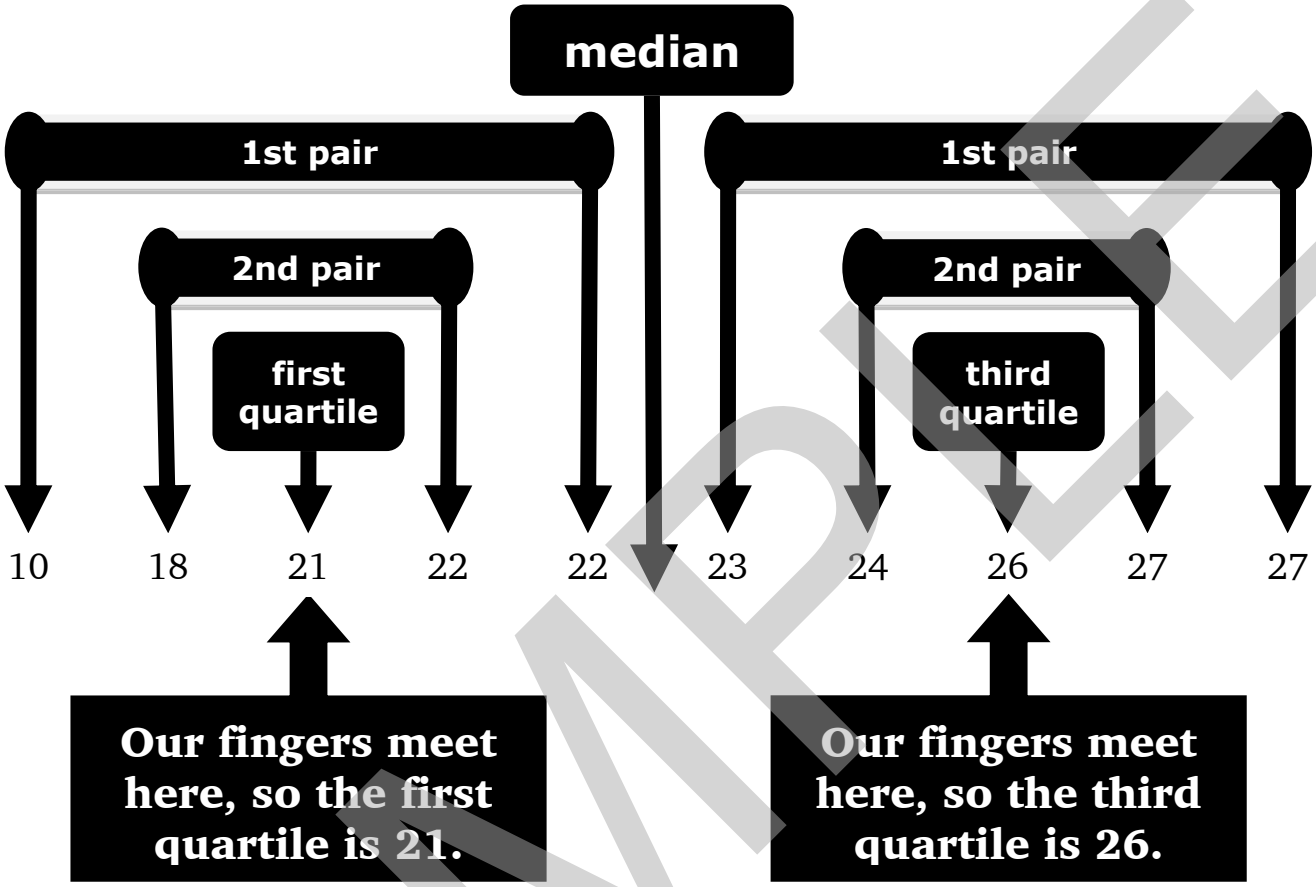
Use your fingers to get the first quartile (the median of this lower half of the data set). **See Figure 5 below.**

Use your fingers to get the third quartile (the median of this upper half of the data set). **See Figure 5 below.**

Our fingers meet here, between the 22 and 23, so the median is the average of those two.

$$\text{median} = \frac{22+23}{2} = \mathbf{22.5}$$

FIGURE 5: Finding the Quartiles using the Finger Method



Solution to Question 5(b)

The first, second and third quartiles are 21, 22.5 and 26, respectively.

Of course, if you prefer, you could also use the $\frac{n+1}{2}$ rule to locate the median and then the first and third quartiles. If you do it properly, your answer will be the same as mine.

6. Find, by hand, the mean, variance and standard deviation of this data (show all your work): 6, 12, 9, 8, 5, 14, 2

Solution to Question 6

It is possible an assignment or exam (although unlikely) will insist you compute a mean and/or standard deviation by hand.

Since the formulas for mean, variance and standard deviation are, respectively, $\bar{x} = \frac{\sum x}{n}$, $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$, and $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{s^2}$, we will set up a table with a column labelled "*x*", the data values which we will total to compute $\bar{x}$; a second column labelled "$x_i - \bar{x}$", where we will find the deviation of each data value from the mean; and a third column labelled "$(x_i - \bar{x})^2$" to get the squared deviations ready to be totalled. Note that I am showing the calculations for the first couple of data values in each column just to make sure you understand what I am doing. Don't show the calculations in each cell, just the results as I do for the rest of the data.

	x	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	6	$6 - 8 = -2$	$(-2)^2 = 4$
	12	$12 - 8 = 4$	$(4)^2 = 16$
	9	1	1
	8	0	0
	5	–3	9
	14	6	36
	2	–6	36
TOTALS	$\sum x = 56$ so $\bar{x} = \frac{\sum x}{n} = \frac{56}{7} = 8$	$\sum (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 102$

We total the first column "$\sum x = 56$" in order to compute the mean, $\bar{x}$, which we need to compute the deviations from the mean in the second column.

$$\bar{x} = \frac{\sum x}{n} = \frac{56}{7} = 8$$

Note, although it is unnecessary in our standard deviation calculation to compute the total of the second column, the total of the deviations from the mean should always be 0. This can act as a check on your work, and it indeed is the case for this problem.

The sum total of all the deviations is always 0.

$$\sum deviations = \sum (x_i - \bar{x}) = 0$$

The third column tells us the total of the squared deviations from the mean:

$$\sum (x_i - \bar{x})^2 = 102$$

We can now compute the **variance**, s^2. Recall, there are $n = 7$ pieces of data.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{102}{6} = 17$$

Finally, we can compute the standard deviation, *s*, by square rooting the variance.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{17} = 4.1231$$

The mean is 8, the variance is 17, and the standard deviation is 4.1231.

Use the *stat mode* on your calculator (see Appendix A to learn how to use the *stat mode* if you haven't done so already) to verify these answers are correct.*

* Note, your calculator will tell you the mean and standard deviation once you have input the seven data values. To get the variance, simply square the standard deviation value (press the "x^2" button on your calculator).

7. In order to analyze the overall pattern of a distribution, the three things we should discuss are:

(A) the mean, median and mode.

(B) the interquartile range, range and variance.

(C) the number of peaks, the outliers and the shape of the distribution.

(D) the shape of the distribution, the centre and the spread.

(E) the outliers, the influential observations and the lurking variables.

Solution to Question 7

Clearly, the correct answer is (D).

8. The annual salary (in thousands of dollars) of a random sample of male and female workers in the construction industry is shown below. Construct a back-to-back stemplot for this data and discuss your observations.

Males: 29 32 32 27 46 24 45 50 47 36 35 30 28 88 37 38 52 43

Females: 28 39 29 23 32 29 18 22 38 40 26 17 33

Notice that the left leaves are increasing away from the stem.

Males Leaf	Stem	Females Leaf
	1	78
9874	2	236899
8765220	3	2389
7653	4	0
20	5	
	6	
	7	
8	8	

A back-to-back stemplot allows us to compare two distributions by using one set of stems for both samples. One sample's leaves are placed to the right of the stem, as usual, while the other's are placed to the left of the stem. Be sure to label each side. You will note in my back-to-back stemplot above, I have collected the

"Males" leaves on the left side of the stems and the "Females" leaves on the right side of the stems. The leaves are placed in *increasing order away from the stem*.

How do we "discuss" our observations?

Words like "discuss" usually make students panic. **The only advice I can offer is first be sure to hit on the key statistical stuff (for example, to "discuss" a distribution be sure to discuss the shape, centre and spread), then use your common sense to translate that information into more practical things.** Here we are comparing the salaries of males and females so tell people who makes more (if anyone). The centre of each distribution can help here. Spread gives us an idea of variety (if a spread were very small that would suggest everyone tends to make the same amount of money; if there was a large spread, there would be a lot of variety in what people make). The shape can also tell us things. For example, if a salary distribution were strongly right-skewed, that would suggest that most people make ordinary amounts of money (where the peak would be), but that long stretch out to the right would say some people make lots and lots of money. If a distribution had two separate peaks, that might suggest there is almost an hierarchy in the salaries (maybe an "entry level" group where the first peak is, and then a second peak where people seem to have graduated to a higher income level). The key is to just say whatever you can think of, and don't worry too much.

For this problem, we must first decide on the shape of the two distributions. Make sure you rotate the stemplot counterclockwise so that the scale is in increasing order when you determine direction of skew (if any). Are one or both of the males and females salary distributions basically symmetrical (ignoring the one outlier of 88 for the males)? Or, are one or both of these distributions slightly right-skewed? I think it is debatable. It is not unusual, especially in assignment questions, to have trouble deciding on the shape of a distribution.

My rule of thumb for deciding skew when looking at a stemplot or histogram is: I want one side of the peak to be at least 2 sections longer than the other side (at least 2 extra rectangles on one side compared to the other, or 2 extra stems) before I start considering a distribution skewed.

If you just can't decide if a distribution is symmetric or skewed, and if time allows, quickly work out the mean and median of the sample (maybe you are asked to do so in another

part of the question anyway). If these two values are approximately the same, that suggests the distribution is symmetric. If they differ significantly, then there must be a skew where the mean was pulled in the direction of the skew. Note that we ignore outliers when determining skew, so work out the mean and median with and without any outliers to get a feel for the shape. NEVER GO THROUGH ALL THIS WORK ON AN EXAM! You simply don't have the time. On an exam, just suck it up and make a decision as quickly as you can.

Looking at the females (personally, I enjoy looking at females), they obviously peak in the 20's, there is only one stem on the left of the 20's (the 10's) and two stems on the right (the 30's and 40's). That is not really enough to convince me of a right-skew (I would want to see at least two extra stems on the right, not just one extra stem). I fed the data into my calculator in *stat mode* and found the mean for the females is 28.8, while the median would be the 7th ordered value (since there are 13 females), which is 29. The mean and median are essentially the same (in fact the mean is a little to the left of the median, but there is no way the distribution looks left-skewed), confirming my opinion that **the females have a symmetric distribution with one peak (in the 20's).** (Personally, I prefer females to be symmetric, preferably with two peaks, but that is a totally different story.)

The males are a little trickier because of the outlier. That value would certainly pull the mean to the right of the median, but that does not necessarily make the distribution right-skewed. Remember, skewness is a property of the data as a whole, not just one or two unusual values. If we keep all 18 males, we find the mean to be 39.9, while the median would be the 9.5 ordered value (the average of the 9th and 10th values) which is 36.5. Clearly the mean is much bigger than the median, but is this just due to the right outlier? If we discard the outlier (88) temporarily, leaving us with 17 males, the mean is now 37.1 (quite a change, demonstrating means are affected by outliers), while the median would now be the 9th value which is 36 (not much of a change at all, demonstrating medians are resistant to outliers). The mean is still larger than the median. **This helps me decide that the males have a slightly right-skewed distribution with one peak (in the 30's) and certainly have a right outlier at 88.**

(Note that there was really only one extra stem on the right side, ignoring the outlier, which was why I was tempted to say the distribution for the males was symmetric. I still think it is debatable, so I would probably include my mean and median values to back up my

conclusion. That way a prof would have to accept my reasoning, and could not mark me wrong. The key is to always justify your conclusions when you yourself have difficulty deciding. If you believe a conclusion about shape is obvious, then they probably think so, too, so no real work is needed to justify it. Also note, when I ignored the outlier to find the mean and median, that was just to help clarify skewness without the outlier's effect. As I said earlier, if a question wants you to compute the mean, median, standard deviation, whatever, never discard data. It is understood they want the result using all the data, outliers and all.)

As far as discussing centre and spread, do not feel obliged to compute means or medians, quartiles or standard deviations, etc. If they have not asked for those things, then they will accept simply rough measures of centre and spread. Of course, you would not be wrong to compute these things, but be mindful of your time. If you do compute these things, do not feel obliged to show any work (use your calculator to do it) since they never asked for it anyway.

When asked to "discuss" or "comment" on a distribution, just try to *B.S.* your way through the discussion. Certainly, since we are comparing two groups, be sure to include comparisons about the shape, centre and spread of the distributions of the two groups. Do they have the same shape or are they different? Do they have the same centre, or is one higher than the other? Do they have the same spread, or is one wider than the other?

If you want to get really fancy, you can also throw in a little speculation as to why the data is the way it is (always stressing that it is only speculation; *why* is always a dangerous question to answer). If you haven't got a clue as to why the data is the way it is, or what that means about the data, then don't talk about it. As Joe Friday in *Dragnet* would say, "Just the facts ma'am." (I know most of you haven't got a clue who he is, Google him if you want to find out). First and foremost, stick to the facts: the numbers you have come up with, the trends you see in your graphs, etc. Only after you have stated the facts, and only if they really request it, you can throw a little interpretation of why the results are the way they are.

Here, then, is how I would "discuss" these distributions:

Solution to Question 8

Males Leaf	Stem	**Females** Leaf
	1	78
9874	2	236899
8765220	3	2389
7653	4	0
20	5	
	6	
	7	
8	8	

The distribution of the female salaries is approximately symmetric with one peak in the 20's, while the male salaries are almost symmetric, but perhaps slightly right-skewed with one outlier at 88 and with one peak in the 30's. (Even ignoring the outlier, the mean and median of the 17 remaining males are 37.1 and 36, respectively, confirming a slight right-skew.) Therefore, the typical female salary tends to be in the 20's (the median is actually 29 thousand dollars) while the males typically earn in the 30's (the median is 36.5 thousand dollars), suggesting that males tend to make more than females. This is further backed up by the fact that the lowest male salary is 24 thousand while the lowest female salary is 17 thousand. Several males make salaries in the 40's and 50's (one even made 88 thousand), while the highest paid female made just 40 exactly. The spread in salaries tends to be about the same (ignoring the outlier for the males), it is just that the males seem to make more money in this industry.

It is not possible for us to say *why* this is so. Is there discrimination towards females? Are females new to the industry and making lower wages due to less experience? (We have no information as to how many years the people in our sample have been working.) These would be questions worth answering through further investigation.

9. The five-number summary for a sample of 60 observations is 27, 45, 50, 62, 101. We can say:

(A) The sample is clearly symmetrical.

(B) The mean is 50.

(C) There are no outliers.

(D) Any data values below 19.5 or above 87.5 are outliers.

(E) Any data values below 24.5 or above 75.5 are outliers.

Note, if you skim through the choices, we see most of them are talking about outliers. That's a hint. The given five-number summary includes the quartiles, so use the **1.5 × *IQR* rule** to establish what values would be considered outliers.

From the five-number summary, we see $Q_1 = \mathbf{45}$ and $Q_3 = \mathbf{62}$. Therefore:

$$IQR = Q_3 - Q_1 = 62 - 45 = \mathbf{17}$$

$$1.5 \times IQR = 1.5 \times 17 = \mathbf{25.5}$$

$$45 - 25.5 = \mathbf{19.5} \text{ and } 62 + 25.5 = \mathbf{87.5}$$

Solution to Question 9

Any values below 19.5 or above 87.5 would be outliers.
The correct answer is (D).

By the way, (A) is clearly wrong. Visualize the boxplot you could make with this five-number summary. Clearly, 50, the given median is not in the centre of the box extending from 45 to 62. The right side of the box is longer (there is a greater distance between 50 and 62, than there is between 45 and 50), so this distribution is right-skewed, not symmetrical.

That also guarantees us (B) is wrong. **The median is 50**, not the mean. And, since the distribution is right-skewed, we would expect the mean to be pulled to the right of the median (the mean will be larger than the median).

Even though we don't know all the data, we are certain (C) is wrong in saying there are no outliers. We showed any data above 87.5 is an outlier. **The maximum is 101**, which is way above 87.5, so it is clearly an outlier. There may be other outliers, but we can't know without seeing the 60 observations.

10. The first and third quartiles for a random sample of 200 observations are 48 and 77, respectively. Three of the observations are 3, 120, and 121. Consider these statements:

(I) 3 is an outlier.

(II) 120 is not an outlier.

(III) 121 is an outlier.

(A) Only (I) is true.

(B) Only (II) is true.

(C) Only (III) is true.

(D) Only (I) and (III) are true.

(E) (I), (II) and (III) are all true.

Again, the **1.5 × *IQR* rule** can be used to determine which of these three numbers, if any, are outliers.

We are given $\boldsymbol{Q_1 = 48}$ and $\boldsymbol{Q_3 = 77}$. Therefore:

$$IQR = Q_3 - Q_1 = 77 - 48 = \mathbf{29}$$

$$1.5 \times IQR = 1.5 \times 29 = \mathbf{43.5}$$

$$48 - 43.5 = \mathbf{4.5} \text{ and } 77 + 43.5 = \mathbf{120.5}$$

Solution to Question 10

Any values below 4.5 or above 120.5 would be outliers.

"3" is an outlier; (I) is TRUE.
"120" is <u>not</u> an outlier; (II) is TRUE.
"121" is an outlier; (III) is TRUE.

The correct answer is (E).

11. In measuring the effectiveness of a new drug treatment for cancer patients, 11 patients were tracked after the treatment to see if the cancer returned. The number of years they were cancer free was recorded, and they were given "*N*" for no cancer if they went at least 10 years without a recurrence. The data was: 3.7 9.8 5.5 *N* 1.2 *N* 7.8 8.8 7.5 *N* 2.9. The median of the data is

(A) 6.5 (B) 7.8 (C) 7.5 (D) *N* (E) impossible to determine

Even though not all of the data is a number, we can still find the median since it is clear that "*N*" is the largest value (representing at least 10 years). **Arrange the data in ascending order and use the Finger Method to find the median.**[*] Working by pairs, counting from the outside in, my fingers meet at "7.8", so that is the median.

median

1.2 2.9 3.7 5.5 7.5 7.8 8.8 9.8 *N* *N* *N*.

Solution to Question 11

The median is 7.8. The correct answer is (B).

If we had been asked to find the <u>mean</u> of this data, that would not have been possible. We can only find the mean if <u>all</u> the data is numerical; whereas, a median is more flexible (all we need is a clear order from smallest to largest).

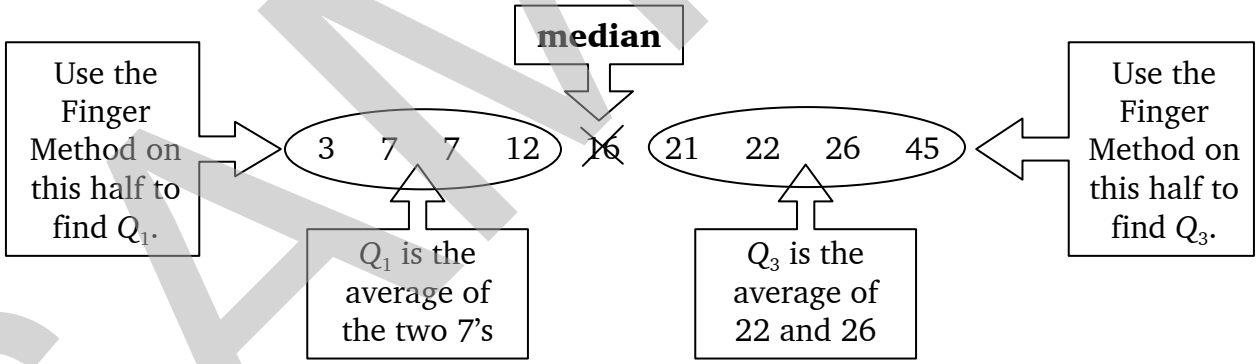
[*] You could also use the $\frac{n+1}{2}$ rule where $\boldsymbol{n = 11}$ if you prefer. We find $\frac{n+1}{2} = \frac{12}{2} = 6$, telling us the median is the 6th ordered data value.

12. The times (in seconds) for 9 subjects to complete a task were as follows: 16 7 3 21 12 7 26 22 45. The interquartile range and variance, respectively, are equal to:
(A) 19; 12.9 (B) 17; 12.9 (C) 15; 165.5 (D) 17; 165.5 (E) 19; 165.5

The variance is simply the square of the standard deviation (it is another measure of spread). The standard deviation of a sample is denoted s (as we mentioned earlier); **the variance of a sample is denoted s^2**. Remember, we can put our calculator into *stat mode*, input the data, and have it tell us what s equals. **ONLY MASOCHISTS USE THE s FORMULA!** See Appendix A if you still don't know how to use *stat mode* on your calculator!

Feeding the data into my calculator, I find $s = 12.8646\ldots$ is the standard deviation; thus, $s^2 = (12.8646\ldots)^2 = 165.5$ (pressing the "x^2" button on the calculator while $s = 12.8646\ldots$ is still on the screen). The **variance is 165.5**, eliminating (A) and (B) as choices.

To get the interquartile range (*IQR*), we must first get the quartiles. I would use the Finger Method since there isn't much data.*



Working by pairs from the outside in, my fingers meet at the "16", so that is the median. My cleaver would smash that value to smithereens, leaving only four numbers in each half. Working by pairs from the outside in for the lower half, my fingers meet between the two 7s, so

* If you prefer, you could use the $\frac{n+1}{2}$ approach. For this problem, $\boldsymbol{n = 9}$, so $\frac{n+1}{2} = \frac{10}{2} = 5$. The median is the 5th ordered data value. **This means that there are 4 pieces of data in front of the median** (as well as 4 values after the median), so let $n = 4$ to find the quartiles: $\frac{n+1}{2} = \frac{5}{2} = 2.5$. The quartiles are the average of the 2nd and 3rd values counted from each end of the ordered data values.

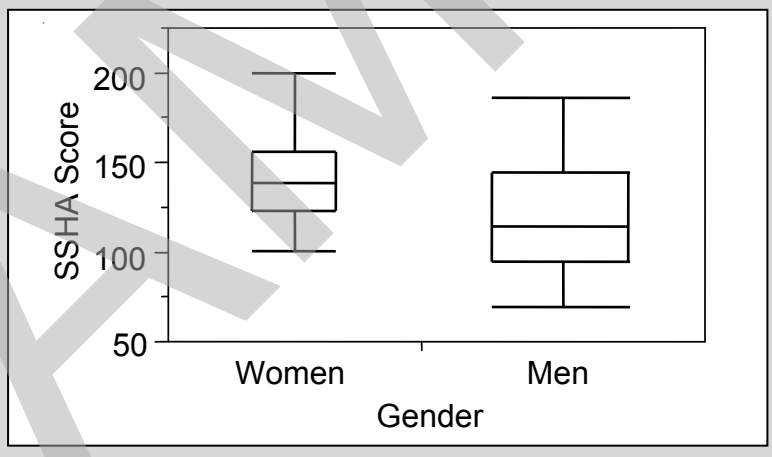
I average them to get the first quartile. Similarly, my fingers meet between the "22" and "26", so I average them to get the third quartile.

Thus: $Q_1 = \frac{7+7}{2} = \mathbf{7}$ and $Q_3 = \frac{22+26}{2} = \mathbf{24}$. Therefore, $IQR = 24 - 7 = \mathbf{17}$.

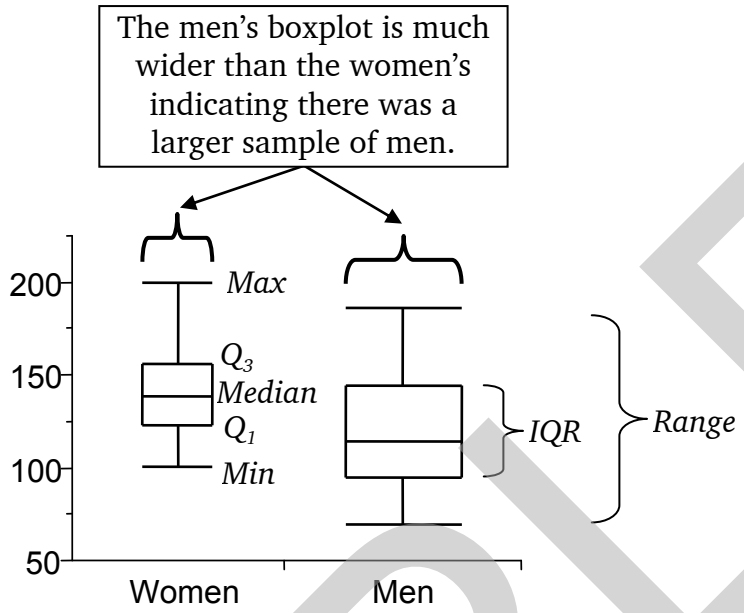
Solution to Question 12

The interquartile range is 17, and the variance is 165.5
The correct answer is (D).

13. The survey of study habits and attitudes (SSHA) is a psychological test that measures the motivation, attitude toward school, and study habits of students. Scores range from 0 to 200. *JMP*™ produced side-by-side boxplots below to compare the scores on the test for a random sample of men and women at the University of Manitoba. Which statement below is <u>false</u>?



(A) The women's scores have a narrower spread than the men's.
(B) The interquartile range for the men's scores is approximately 50.
(C) More men wrote this test than women.
(D) The women tend to score higher than the men on this test.
(E) The mean test score for the men is less than their median score.



Side-by-side boxplots are commonly used to compare two or more sample distributions using the same scale. Here we can easily see the similarities and differences between men and women's SSHA test scores.

(A) is TRUE. Both the distance from max to min (the range) and the length of the box (the interquartile range) for the women is smaller than the men's.

(B) is TRUE. The length of the men's box, and so the *IQR*, is approximately 50.

(C) is TRUE. This is just a stupid fact we have to know about how *JMP*™ draws boxplots. The wider the boxplot, the larger the sample it was based on. We can clearly see that the men have a much wider boxplot than the women, so there must have been more men in the sample. Do not think you would ever have to do something like this if you were drawing two boxplots yourself.

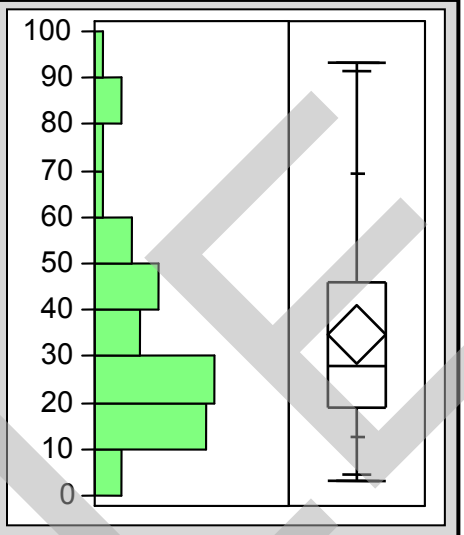
(D) is TRUE. Each value of the five-number summary is higher for the women than the men.

(E) is FALSE. The right side of the box (between the median and Q_3) is longer than the left side, and the right whisker is longer than the left whisker, suggesting the men have a right-skewed distribution. The mean will be pulled to the right of the median, making it larger, not smaller.

Solution to Question 13

The correct answer is (E).

14. At right is the distribution of annual income (in thousands of dollars) of a sample of Canadian adult males as displayed by *JMP*™.



(a) From this information we would conclude:

(A) the sample was clearly not random.

(B) the distribution is symmetrical.

(C) the distribution is left-skewed.

(D) the distribution is right-skewed.

(E) men make more money than women.

(b) The best way to summarize the above data is:

(A) the mean and standard deviation.

(B) the mean and variance.

(C) the mean, the median and the mode.

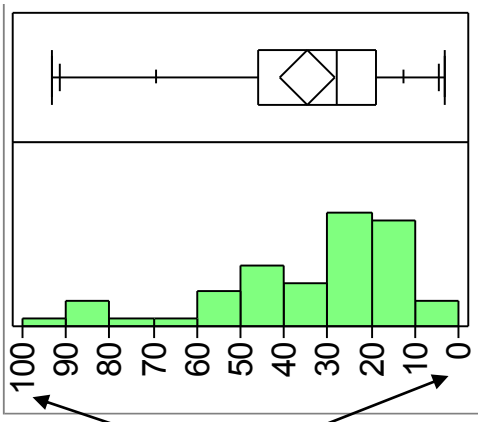
(D) the five-number summary.

(E) none of the above.

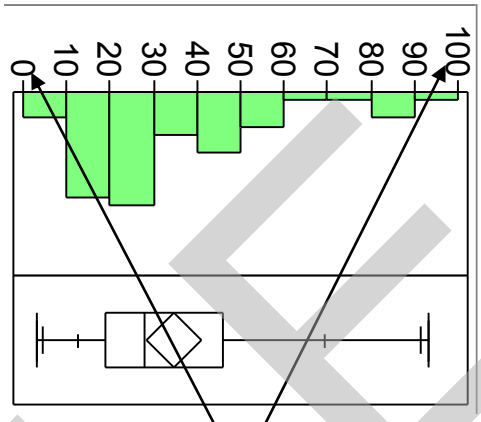
(c) The mean and median, respectively, of this distribution are approximately

(A) 28; 35 (B) 35; 28 (C) 45; 32 (D) 32; 45 (E) 40; 32

Clearly, the distribution is skewed (we see this in both the histogram and the box of the boxplot). If we rotate the graphs properly (make sure that the scale moves from small numbers on the left to large numbers on the right) and cut the histogram at the peak in the "20-30" class, the right side (the larger number side) is clearly much longer than the left side. This is also confirmed by the box in the boxplot where the side of the box to the right of the median line is longer and the right whisker is longer than the left whisker (we can trust the whiskers since there is apparently no outlier according to the histogram).



If you rotate the picture this way you would be fooled into thinking the distribution is left-skewed. WRONG! Your scale is the wrong way round; "0" should be on the left and "100"



This is the way to rotate the picture to see the direction of skew. The scale is now running from "0" on the left to "100" on the right as it should.

Solution to Question 14(a)

There is no doubt that the distribution is right-skewed.
The correct answer is (D).

14. (b) The best way to summarize the above data is:

(A) the mean and standard deviation.

(B) the mean and variance.

(C) the mean, the median and the mode.

(D) the five-number summary.

(E) none of the above.

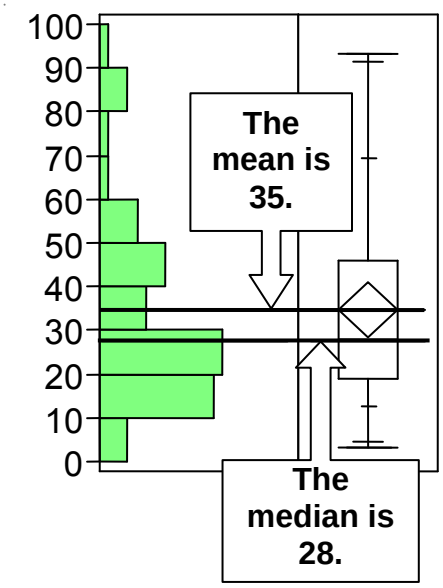
Since the distribution is skewed, it would be inappropriate to compute the mean as a measure of centre. We should compute the five-number summary since this contains the median and quartiles which are *resistant to skewness*. The five-number summary will give us a fair depiction of the shape, centre and spread of the distribution.

Solution to Question 14(b)

The correct answer is (D).

14. (c) The mean and median, respectively, of this distribution are approximately

(A) 28; 35 (B) 35; 28 (C) 45; 32 (D) 32; 45 (E) 40; 32



Since the distribution is right-skewed, we know the mean will be pulled to the *right* of the median. Therefore, the mean is certainly larger than the median, eliminating (A) and (D) as choices. We can more accurately estimate the median since we know it is the line inside the box of the boxplot. That line is approximately "28" on the scale. Similarly, *JMP*™ included a diamond in this box. That is called a "Mean Diamond". The centre of the diamond denotes the position of the mean. This is approximately "35" on the scale. **Therefore, the mean is 35 and the median is 28.**

Solution to Question 14(c)

The correct answer is (B).

What if we only had the histogram to look at?

In a single-peaked symmetrical distribution, the mean, median and mode would all be the same. In those cases, you would then estimate the value of all three to be the centre of the peak. In a single-peaked skewed distribution, the skew certainly pulls the mean away from the peak (in the direction of the skew); we also assume the median is pulled slightly away from the peak (in the direction of skew also), the mode is assumed to be at the centre of the peak, of course. Which is to say, if a distribution is left-skewed, the mode stands at the centre of the peak, the median is pulled a little to the left of the peak, and the mean is pulled even further to the left of the peak. I would assume that the median is still in the same rectangle as the mode, just closer to the left side of the rectangle. I would assume the mean has been pulled far

enough left that it would be in the rectangle next to the peak, maybe about the centre of that rectangle. Obviously, this is reversed if the distribution is right-skewed.

For example, in this problem, the peak is clearly the "20 – 30" class. I would assume that the mode is at the centre of the peak (the mode = 25). Since there is a right-skew, the median has probably been pulled a little to the right of the mode, but still sits in the same class. I would estimate the median to be somewhere between 25 and 30 then, like 27 or 28 (which, of course, nicely fits the correct answer of 28). The right-skew will have a stronger pull on the mean, so I would assume it is pulled right, landing in the next rectangle (the "30 – 40" class) about halfway. I would estimate the mean to be about 35. (Which is exactly what the correct answer is! Amazing isn't it?) Again, these are only estimates, and there are exceptions, but they should give you a pretty good idea what the correct choice would be on a multiple-choice test.

15. The table below shows the scores for the 60 people who participated in a recent 18-hole golf tournament (the winner scored −1 or 1 below par).

Score	**−1**	**0**	**+1**	**+2**	**+3**	**+4**	**+5**	**+6**	**+7**
Frequency	**1**	**0**	**1**	**4**	**3**	**7**	**10**	**25**	**9**

Looking at the distribution of scores, we can conclude:

(A) The median is +3 and the mean is higher.

(B) The median is +3 and the mean is lower.

(C) The median and mean are both +3.

(D) The median is +6 and the mean is lower.

(E) The median and mean are both +6.

Looking at the frequency table we are given, we can visualize the histogram we could make for it. We would have a little bar at "−1" (since it happened only once), no bar at "0", a little bar at "+1", and slightly higher bars at "+2" and "+3" (since they have frequencies of 4 and 3, respectively). The bars then start increasing quite a bit peaking at "+6", where the frequency is 25, before dropping down at "+7". **Clearly, the distribution is single-peaked (at "+6") and skewed left (the left side of the peak stretches all the way back to "−1" while there is only "+7" on the right side of the peak).**

The sample size is $\boldsymbol{n = 60}$, so $\frac{n+1}{2} = \frac{61}{2} = 30.5$. That means the median is between the 30th and 31st ordered score. The frequency table has already put the scores in order for us. The last thirty-four scores are the twenty-five "+6" scores and the nine "+7" scores so, if you were counting from the right end, the 30.5 mark would land between two of the "+6" scores. Therefore, **the median is +6**. By the way, the mode is also +6 since that is far and away the most frequent score.

We have already noticed the distribution is left-skewed, so the mean is pulled to the left of the median. **The mean is lower than the median.**[*]

Solution to Question 15

The median is +6 and the mean is lower. The correct answer is (D).

16. The stemplot at right shows the number of weeks a sample of 40 patients in Manitoba had to wait before receiving hip replacement surgery. The five-number summary for this data is:

(A) 7, 26, 35, 42, 52

(B) 0, 26.5, 35, 42.5, 69

(C) 7, 26, 35, 42, 69

(D) 7, 26.5, 35, 42.5, 69

(E) 7, 26.5, 35, 40.5, 69

Stem	Leaf
7	
6	9
6	
5	
5	12
4	67789
4	0011234
3	555589
3	13344
2	55667789
2	2
1	589
1	4
0	7

[*] If you are having trouble visualizing all this think of the sixty scores lined up in order: −1, +1, +2, +2, +2, +2, +3, +3, +3, … +7, +7, +7, +7, +7, +7, +7, +7, +7. No matter which end you count from, the 30.5 position will put you in the same place: between two of the +6 scores. If you don't believe me when I conclude the mean will be lower than the median due to the left-skew, go ahead and compute it:

$\bar{x} = \frac{\sum x}{n} = \frac{-1+(+1)+(4\times+2)+(3\times+3)+(7\times+4)+(10\times+5)+(25\times+6)+(9\times+7)}{60} = \frac{308}{60} = 5.13$, lower than +6.

Don't worry that the stems are descending instead of ascending. That's just *JMP*™ up to its tricks again.

Stem	Leaf
3	5 5 5589 20 21
3	1 3 3 4 4 15 16 17 18 19
2	5 5 6 6 7 7 8 7 8 9 10 11 12 13
	9 14
2	2 6
1	5 8 9 3 4 5
1	4 2
0	7 1

Clearly, the minimum value is 7 ("07") and the maximum value is 69 (don't be distracted by the "7" stem which has no leaves), eliminating (A) and (B) as choices.

Since $\boldsymbol{n = 40}$, $\frac{n+1}{2} = \frac{41}{2} = 20.5$, telling us the median is the average of the 20th and 21st ordered value. As the count at right shows (count the leaves, <u>not</u> the stems!), this takes us between the first two 35s. **The median equals** $\frac{35+35}{2} = \mathbf{35}$. None of the remaining choices can be eliminated.

The median is at the 20.5 position, so there are 20 pieces of data before and after the median. We use $n = 20$ to find our quartiles: $\frac{n+1}{2} = \frac{21}{2} = 10.5$. The first and third quartile are each the average of the 10th and 11th values in their respective halves. Again, from the count above, we see that Q_1 is the average of 26 and 27. $\boldsymbol{Q_1 = \frac{26+27}{2} = 26.5}$. This eliminates (C) as a possible choice.

Stem	Leaf
6	9 ← 1
6	
5	
5	1 2 ← 3 2
4	6 7 7 8 9 ← 8 7 6 5 4
4	0011 2 3 4 ← 11 10 9

Be careful when counting your way to Q_3! Start from the maximum value and count from the *max* of each stem, as we see at right. As we have said before, count in from the Min to find Q_1 and count in from the Max to find Q_3. **In the case of a stemplot, count from the Min of <u>each</u> stem on your way to Q_1 and count from the Max of <u>each</u> stem on your way to Q_3.** Look very closely at the way I am counting in the diagram at right. The third quartile is, therefore, the average of 42 and 43. $\boldsymbol{Q_3 = \frac{42+43}{2} = 42.5}$.

Solution to Question 16

The 5-number summary is 7, 26.5, 35, 42.5, 69. The correct answer is (D).

THE EFFECT OF CHANGING UNITS ON CENTRE AND SPREAD

Consider a situation where you collected a random sample of 1,000 12-year old boys and recorded their height, and reported **the five-number summary (in inches) is 44, 48, 52, 58, 69**. Here, we see the median height is 52 inches, while the minimum and maximum heights were 44 and 69 inches, respectively. Now, you want to report the information in metric (centimetres to be precise). You know 1 inch = 2.54 cm, so it is merely a matter of multiplying every height by 2.54 to convert it to centimetres. If we let Cm = height in centimetres, and I = height in inches, we could write the conversion formula as:

$$\boldsymbol{Cm = 2.54I}$$

To report the five-number summary in centimetres, we don't have to go and convert all 1,000 pieces of data and then redo the summary. Clearly, it is merely a matter of converting the five-number summary itself. For example, the shortest boy was 44 inches tall, that converts to 44 × 2.54 = 111.76 cm tall. That is still obviously going to be the minimum height. The taller boys will obviously be taller in centimetres, too.

Thus, multiplying the five numbers by 2.54, we see **the five-number summary (in centimetres) is 111.76, 121.92, 132.08, 147.32, 175.26.**

We know the median is a measure of centre; note how the median converted from 52 inches to 132.08 centimetres (i.e. **all we had to do is multiply the median in inches by 2.54 to get the median in centimetres**). The range, a measure of spread, used to be 69 – 44 = 25 inches. Now, the range is 175.26 – 111.76 = 63.5 centimetres. **The range in centimetres is the range in inches multiplied by 2.54.** Check: 25 × 2.54 = 63.5. Similarly, the interquartile range (another measure of spread) in inches is 58 – 48 = 10 inches. In centimetres, the interquartile range is 147.32 – 121.92 = 25.4 centimetres. Again, **the interquartile range in centimetres is simply the *IQR* in inches multiplied by 2.54** (10 × 2.54 = 25.4).

Let's say we had found the mean and standard deviation of the 12-year old boys to be 52.7 inches and 7.35 inches, respectively. Then, it is merely a matter of multiplying these values by 2.54 to get the corresponding values in centimetres. We find the mean height is 133.858 cm with a standard deviation of 18.669 cm.

If you are converting your data into new units by simply multiplying the old values by some constant (like 2.54, for example). Then any measure of centre and spread will also be multiplied by that constant. There is no need to convert all the data into the new units, and then recompute the measures of centre or spread. Instead, simply multiply the old measures by the conversion constant.

Now consider a set of data compiled from a random sample (the size is irrelevant) of vendors at a local farmers market. Each vendor reported their gross sales for the day in dollars. You report **the gross sales five-number summary is 12, 45, 70, 110, 175 dollars. Thus, the median is \$70, the range is \$163 and the *IQR* is \$65** (computing *Max* – *Min* and $Q_3 - Q_1$, respectively). **The mean and standard deviation are found to be \$78.75 and \$44.26, respectively.**

Now, you are interested in determining the profits for each vendor. Each vendor had to pay \$25 for their place at the market, and let's pretend there were no other costs involved. (I know that is ridiculous, but give me some slack! I am trying to show you something here, without making things too complicated.) Thus, we simply have to subtract \$25 from each vendor's gross sales to determine their profit for the day. If we let S = gross sales, and P = profits, a mathematician might write the conversion formula:

$$\boldsymbol{P = S - 25}$$

Subtracting 25 from each value, **the profits five-number summary is –13, 20, 45, 85, 150 dollars. The median profit is \$45, simply the median sales minus 25, as expected.** Here is where things get interesting: **The range of the profits is 150 – (–13) = \$163, and the *IQR* = 85 – 20 = \$65. The range and interquartile range are exactly the same for profits as they were for gross sales!**

If you are converting your data by subtracting (or adding) a constant (like subtracting 25, for example), then you will do likewise to convert any measure of centre. MEASURES OF SPREAD ARE COMPLETELY UNAFFECTED BY ADDING OR SUBTRACTING A CONSTANT IN A CONVERSION FORMULA.

Thus, since the mean sales is \$78.75, **the mean profits are 78.75 – 25 = \$53.75**. However, **the standard deviation for the profits will be exactly the same as it was for the sales, \$44.26,** since that is a measure of spread.

If we are converting data into new units via a conversion formula of some sort, then the measures of centre and spread will convert also according to these principles:

Multiplying (or dividing) by a constant will change all measures of centre and spread accordingly. Adding (or subtracting) a constant will only affect measures of centre (such as mean or median); adding (or subtracting) will have no effect on measures of spread (such as standard deviation or interquartile range).

Which is to say, if we are converting data from *X* units into *Y* units via the conversion formula

$$Y = AX + B$$

where *A* and *B* are constants, then:

$$\textit{Centre of } Y = A(\textit{Centre of } X) + B$$

Apply the entire formula to measures of centre in *X* units (such as mean, median or mode) to get measures of centre in *Y* units.

$$\textit{Spread of } Y = A(\textit{Spread of } X)$$

Only multiply measures of spread in *X* units (such as standard deviation, range or *IQR*) by the constant *A* to get measures of spread in *Y* units; the "addition constant" *B* is irrelevant to measures of spread.

For example, if we are told $Y = 13 + 5X$, and we know the mean and standard deviation of X is 10 and 3, respectively, then the mean of $Y = 13 + 5 \times 10 = 63$, but the standard deviation of Y is simply $5 \times 3 = 15$. We do not add the 13 to the standard deviation!

17. The mean daily high temperature for June in a particular city is 20 °C with a standard deviation of 2.7 °C. What would the mean and standard deviation be in Fahrenheit? (Hint: The formula $F = 1.8C + 32$ converts Celsius into Fahrenheit.)

(A) 68 & 3.86 (B) 68 & 36.86 (C) 68 & 4.86 (D) 68 & 34.7

(E) It is impossible to determine without the original data.

We are given the **mean is 20 °C**. The mean is a measure of centre. Simply apply the conversion formula to the mean in Celsius to get the mean in Fahrenheit.

$$F = 1.8 \times 20 + 32 = \mathbf{68}$$

The mean is 68 °F. The correct answer could be any of (A) through (D).

We are given the **standard deviation is 2.7 °C**. The standard deviation is a measure of spread. **Only multiplication or division affects spread.** The conversion formula is $F = 1.8C + 32$, so the "32" will have no effect on spread (it is an addition constant). Only the "1.8" multiplying constant is of consequence.

The standard deviation $= 1.8 \times 2.7 =$ **4.86 °F**.

Solution to Question 17

The correct answer is (C).

Put another way, since the formula is $F = 1.8C + 32$, we can say:

$$\textbf{\textit{Centre of }}F = 1.8(\textbf{\textit{Centre of }}C) + 32$$

$$\textbf{\textit{Spread of }}F = 1.8(\textbf{\textit{Spread of }}C)$$

Thus, since mean is a measure of centre while standard deviation is a measure of spread:

$$\textbf{\textit{Mean of }}F = 1.8(\textbf{\textit{Mean of }}C) + 32$$

$$\textbf{\textit{Standard Deviation of }}F = 1.8(\textbf{\textit{Standard Deviation of }}C)$$

18. After a particularly difficult exam, the average mark was 47%, the median was 40%, and the interquartile range was 10%. The prof decides to add 20% to everyone's mark. After this addition, which of the following statements is FALSE?

(A) The first quartile could now be as low as 50%.

(B) The third quartile could now be as high as 70%.

(C) The median mark is now 60%.

(D) The interquartile range is now 30%.

(E) The average mark is now 67%.

Since the prof is adding 20% to every mark, any measure of centre will also have 20% added to it. Thus, **the average mark will increase from 47% to 67% ((E) is TRUE)**, and **the median mark will increase from 40% to 60% ((C) is TRUE)**.

But, addition (or subtraction) has no effect on spread. So, the interquartile range will remain at 10%. (D) is FALSE!

Solution to Question 18

The correct answer is (D).

By the way, obviously (A) and (B) must be true. Think about it. We know the median is now 60%, and, since the *IQR* is 10%, there must be a 10% spread between Q_1 and Q_3. **There is no law that says the first, second and third quartiles have to all be distinctly different numbers.** For instance, if everyone in the class got the same mark (let's say 60%), then you would be lining up a bunch of 60s all in a row. Obviously, then, the five-number summary would be 60, 60, 60, 60, 60 (the minimum is 60, the maximum is 60, and so is everything in between). Imagine the boxplot for that! It would end up being just one fence at 60 (since all five fences would be superimposed on top of each other). This, of course, is not the case in this problem since we know $IQR = 10$, whereas, in the scenario I just outlined, *IQR* would be 0.

Obviously, Q_3 could not be any lower than 60% (because the median is 60%, and Q_3 could not be lower than the median; you can't reach the third quartile before you reach the

median, which is the second quartile). Therefore, the lowest Q_1 could be is 50% (since IQR = 10%). It is conceivable to have $Q_1 = 50$, $Median = 60$, and $Q_3 = 60$. Similarly, Q_1 could not be any higher than 60% (because the median is 60%, and you can't reach the second quartile before you reach the first quartile). Therefore, the highest Q_3 could be is 70%. It is conceivable to have $Q_1 = 60$, $Median = 60$, and $Q_3 = 70$. Put another way, **the median can be as low as the first quartile or as high as the third quartile.**

19. A student's average after writing six tests is 68%. She just got her seventh test back with a mark of 82%. What is her average mark now?
(A) 74% (B) 71% (C) 75% (D) 72% (E) 70%

The mistake many students will make in a question like this is to simply average the 68 and 82: $\frac{68+82}{2}=75$. They then think the correct answer is 75%. WRONG! You must take into account that the 68% was based on 6 tests. That deserves much more weight than the single test mark of 82%.

The key in any question that gives you a mean but then proceeds to change the data is to establish the total of the data. Recall: when you are given the mean of a sample, then, for purposes of argument, it is as if every member of the sample has that mean score. In this problem, we were told the mean for the 6 tests is 68%, so that is as if she scored 68 on every single one of those tests. Visualize that. It is like we have a column recording all her test scores, and, in that column, we have entered 68 all six times. Now, she has written a seventh test and scored 82, so we have an 82 entered in the seventh column (see the table at right). Add all these scores up and we come up with a total of 490 (68 × 6 + 82 = 490). We can now compute her new average: $\bar{x}=\frac{\sum x}{n}=\frac{490}{7}=\mathbf{70}$.*

Test	Test Score
1	68
2	68
3	68
4	68
5	68
6	68
7	82
Total	490

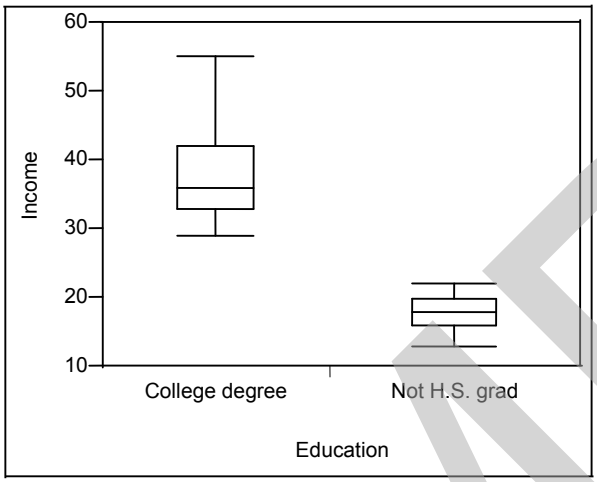
Solution to Question 19

Her average mark is now 70%. The correct answer is (E).

* Alternatively, you could put your calculator into *stat mode* and enter these seven scores to find the mean.

20. Below is a *JMP*™ printout comparing the annual income (in thousands of dollars) of two random samples of 35-year-old full-time workers. One sample is of 10 people who have a college degree, the other is 11 people who did not graduate from high school.

Oneway Analysis of Income By Education



Quantiles

Level	Minimum	10%	25%	Median	75%	90%	Maximum
College degree	29	29.1	33	36	42	54.3	55
Not H.S. grad	13	13.2	16	18	20	21.8	22

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
College degree	10	38.1000	7.97844	2.5230	32.393	43.807
Not H.S. grad	11	18.0000	2.82843	0.8528	16.100	19.900

What is the mean and median income of all 21 workers?

(A) $27,571 and $27,000, respectively

(B) $27,571 and $22,000, respectively

(C) $27.571 and $27, respectively

(D) $28,050 and $27,000, respectively

(E) $28,050 and $22,000, respectively

This question is similar to the last one. We don't want to make the mistake of averaging the two means to get the overall mean. Again, to find a new mean anytime they change the data in some way, we first find the total of all the data.

We can see in the "Means and Std Deviations" table, the 10 college degree workers have a mean of 38.1 ($\boldsymbol{n_1 = 10}$ and $\boldsymbol{\bar{x}_1 = 38.1}$). So, for purposes of argument, we can visualize a table of data where the first 10 workers all have a salary of 38.1 thousand dollars. We also see the 11 workers who did not finish high school have a mean of 18 ($\boldsymbol{n_2 = 11}$ and $\boldsymbol{\bar{x}_2 = 18}$). Again, we can visualize the next 11 workers on our table (the 11th to 21st worker in total) all making a salary of 18 thousand dollars. Adding all these salaries up, **the total for all 21 workers is 579 thousand dollars** (38.1 × 10 + 18 × 11 = 579). We can therefore compute the mean for all 21 workers:

Worker	Salary
1	38.1
2	38.1
3	38.1
. . .	. . .
9	38.1
10	38.1
11	18
12	18
13	18
. . .	. . .
20	18
21	18
Total	579

$$\bar{x} = \frac{\sum x}{n} = \frac{579}{21} = \mathbf{27.571}$$

The mean is 27.571 thousand dollars (i.e. $27,571). The correct answer is either (A) or (B) (note (C) is incorrect; it says the mean is 27.571 dollars, not 27.571 thousand dollars).

We still need to determine the median of all 21 workers. Again, do not think you could just average the two medians we were given in the "Quantiles" table. Many students will make the mistake of thinking, since the median of the college degree workers is 36 and the median of those who did not finish high school is 18, they could just average the two values: $\frac{36+18}{2} = 27$ and declare the median is 27 thousand dollars (which is choice (A)). WRONG!

The only way to find a median is to be able to line all the data up from smallest to largest. But, in this question, we have not even been given the original data. There must be a trick to it. Observe, in the "Quantiles" table we have been given the 5-number summary for each group (as well as a couple of other numbers we aren't interested in). **The 5-number summary for the "College degree" sample is 29, 33, 36, 42, 55 while, for the "not H.S. grad" sample, it is 13, 16, 18, 20, 22.** What is really important is that we notice the Maximum value of the "not H.S. grad" workers is 22, while the Minimum of the

"college degree" workers is 29. **All eleven workers who did not finish high school have lower incomes than any of the college degree workers! This is confirmed in the side-by-side boxplots.** If we lined up all 21 values in order from smallest to largest, the first 11 would be the "not H.S. grad" workers, then the last 10 would be the "college degree" workers.

To find the median of $n = 21$ workers, we see $\frac{n+1}{2} = \frac{22}{2} = 11$, telling us the 11th ordered value is the median. Since the lowest income people are the 11 "not H.S. grad" workers, the 11th number would be the very last one of them (the maximum of that group, 22). **The median is 22 thousand dollars ($22,000).**

Solution to Question 20

The mean and median are $27,571 and $22,000, respectively.
The correct answer is (B).

If the two samples had overlapped (for example, if the maximum for the "not H. S. grad" had been 30 and the minimum for the "College degree" sample had been 22) there would have been no way to tell what the median is without knowing the actual data. The best we could do is perhaps use the boxplots to make an educated guess as to approximately where the median might be. **In general, you have to know exactly what the data is to be able to determine the median. If you have been told only what some of the data is, you can bet, on an exam, you have been told enough to determine the median. If all you have is a graph to look at (like a histogram), the best you can do is approximate the median value by approximating what value on the graph would leave 50% of the data on each side.**

SUMMARY OF KEY CONCEPTS IN LESSON 1

- Know the difference between a **quantitative variable** and a **categorical variable**.
  - If the variable is quantitative, is it **continuous** or **discrete**?
  - If the variable is categorical, is it **nominal** or **ordinal**?
- Use a **histogram**, **stemplot**, or **boxplot** to display the distribution of a quantitative variable.
- Use a **bar chart** or **pie chart** to display the distribution of a categorical variable.
- Use a **time series** to display data that has been collected as time goes by in order to identify trends, if any, in that data.
- When making a stemplot, we can choose to make a **split stemplot** to count by fives rather than tens if that will display the data better. We can also **trim the data** (cut away the last digit) if that will make the number of stems more manageable.
- When making a boxplot, we can choose to use a **modified boxplot** (also called an **outlier boxplot**) to display any outliers by dots in order to not exaggerate the length of the whiskers.
- If we want to compare two separate distributions graphically, we can use a **back-to-back stemplot** or **side-by-side boxplots**. Side-by-side boxplots can even be used to compare three or more distributions on one graph.
- Discuss the **shape**, **centre** and **spread** of a quantitative variable when summarizing its distribution.
  - **Shape:** Consulting any graphs you have, are there any **outliers**? How many **peaks** are there, if any? Is the distribution **symmetric**, **left-skewed** or **right-skewed**?
  - **Centre:** Three measures of centre are **mean**, **median** and **mode.** For symmetric data, the mean and median are the same. **When data is skewed or has outliers, the mean is pulled away from the median in the direction of the skew or the outliers. A median is a more trustworthy measure of centre in these situations.**
  - **Spread:** Three measures of spread are **range**, **interquartile range** and **standard deviation**. When using a median to measure the centre, use the range and interquartile range to measure the spread. When using a mean to measure the centre, use the standard deviation to measure the spread.
- The **five-number summary** is ***Min, First Quartile, Median, Third Quartile, Max***.

- Know how to use both **the $\frac{n+1}{2}$ Rule** and **the Finger Method** to locate the median and/or quartiles and use whichever you prefer for a given data set. Remember, use a cleaver to slice the ordered data in half at the median's location. If that cleaver hits a piece of data, it has been smashed to smithereens, excluding it from both the lower and upper halves when it comes to counting your way to the quartiles.
- The **1.5 × *IQR* Rule** can be used to determine outliers. Any value that is lower than $Q_1 - (1.5 \times IQR)$ or higher than $Q_3 + (1.5 \times IQR)$ is an outlier.
- The **sample mean** is denoted $\bar{x}$ and the **sample standard deviation** is denoted s. The **sample variance** is s^2.
- Memorize these formulas: $\bar{x} = \frac{\sum x}{n}$ and $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$.
- If you are ever asked to transform data by adding, subtracting, multiplying, and/or dividing each value by some constant, remember how that transforms any measure of centre or spread. **Any transformation you are performing on the data will also transform the measures of centre in the same way. However, only multiplication or division transforms a measure of spread. Addition or subtraction by a constant has no effect on spread.**
  - If we are transforming X into Y via the formula $Y = AX + B$, then
    - ***Centre of* Y = A (*Centre of* X) + B**

      (The mean, median, or mode would be transformed this way.)
    - ***Spread of* Y = A (*Spread of* X)**

      (We do not add the constant "B"; standard deviation, range, or interquartile range would be transformed this way.)

LECTURE PROBLEMS FOR LESSON 1

For your convenience, here are the 20 questions I used as examples in this lesson. Do not make any marks or notes on these questions below. Especially, do not circle the correct choice in the multiple choice questions. You want to keep these questions untouched, so that you can look back at them without any hints. Instead, make any necessary notes, highlights, etc. in the lecture part above.

1. A survey asked the following questions:

- What is your eye colour?
- Rate your boss on a scale from 1 to 10 where 1 is awful and 10 is wonderful.
- How much do you weigh (in kilograms)?
- What is the 3-digit area code for your home phone number?
- How far do you live from work or school? (less than 5 km, between 5 km and 10 km, more than 10 km)
- What method of transportation to work/school do you normally use? (car, bus, bike, walking, other)
- What is your annual income?
- How many times in a typical month do you eat at a restaurant (including take-out or delivery)?
- Did you vote in the last federal election?

For each of the above survey questions, identify if the variable is quantitative (if so, is it discrete or continuous?) or categorical (if so, is it ordinal or nominal?). In addition, what graph could you use to display the data gathered in each case?

(See the solution on page 14.)

2. You record the age, marital status, earned income, and sex of a sample of 1463 people. The number of variables you have recorded is:

(A) 1463.
(B) five: age, marital status, income, sex, and number of people.
(C) four: age, marital status, income, and sex.
(D) two: age and income; marital status and sex are not variables because they are not a numerical quantity.
(E) none: because no one has any business asking such personal questions.

(See the solution on page 15.)

3. The table below shows the number of people (in millions) living on farms in a certain country over the years.

Year	1910	1920	1930	1940	1950	1960	1970	1980	1990
Population	1.6	2.9	2.8	2.5	1.8	1.3	1.3	0.9	0.6

Construct a time series for this data and comment on what you see.
(See the solution on page 16.)

4. The test scores (out of 100) for a random sample of 50 students who wrote a statistics midterm exam are as follows:

75	88	47	66	78	45	73	66	77	100
64	61	77	87	66	92	86	57	80	70
52	84	80	79	66	92	72	83	50	84
65	75	77	79	79	57	63	51	44	59
84	77	44	81	61	77	57	75	3	52

(a) Construct a frequency table. *(Solution on page 23.)*

(b) Construct a relative frequency table. *(Solution on page 24.)*

(c) Construct a histogram. *(Solution on page 26.)*

(d) Construct a stemplot. *(Solution on page 28.)*

(e) Construct a split stemplot. *(Solution on page 30.)*

(f) Discuss the shape of the distribution. Are there any outliers? *(Solution on page 36.)*

(g) Find the median test score. *(Solution on page 37.)*

(h) Find the first and third quartiles. *(Solution on page 40.)*

(i) Find the Range and Interquartile Range. *(Solution on page 41.)*

(j) State the five-number summary. *(Solution on page 42.)*

(k) Justify the outliers (if any) mathematically. *(Solution on page 43.)*

(l) Draw a boxplot. *(Solution on page 44.)*

(m) Draw a modified (outlier) boxplot. *(Solution on page 46.)*

(n) Find the mode of the distribution. *(Solution on page 47.)*

(o) Find the mean of the distribution. *(Solution on page 49.)*

(p) Find the standard deviation of the distribution. *(Solution on page 51.)*

5. Find the first, second and third quartiles for the data sets below.

(a) 3, 8, 84, 51, 23, 13, 18, 15, 18, 4, 16, 4, 9.

(See the solution on page 55.)

(b) 24, 27, 26, 22, 23, 27, 22, 18, 21, 10.

(See the solution on page 60.)

6. Find, by hand, the mean, variance and standard deviation of this data (show all your work): 6, 12, 9, 8, 5, 14, 2

(See the solution on page 61.)

7. In order to analyze the overall pattern of a distribution, the three things we should discuss are:

(A) the mean, median and mode.

(B) the interquartile range, range and variance.

(C) the number of peaks, the outliers and the shape of the distribution.

(D) the shape of the distribution, the centre and the spread.

(E) the outliers, the influential observations and the lurking variables.

(See the solution on page 63.)

8. The annual salary (in thousands of dollars) of a random sample of male and female workers in the construction industry is shown below. Construct a back-to-back stemplot for this data and discuss your observations.

Males: 29 32 32 27 46 24 45 50 47 36 35 30 28 88 37 38 52 43
Females: 28 39 29 23 32 29 18 22 38 40 26 17 33

(See the solution on page 67.)

9. The five-number summary for a sample of 60 observations is 27, 45, 50, 62, 101. We can say:

(A) The sample is clearly symmetrical.

(B) The mean is 50.

(C) There are no outliers.

(D) Any data values below 19.5 or above 87.5 are outliers.

(E) Any data values below 24.5 or above 75.5 are outliers.

(See the solution on page 68.)

10. The first and third quartiles for a random sample of 200 observations are 48 and 77, respectively. Three of the observations are 3, 120, and 121. Consider these statements:

(I) 3 is an outlier.
(II) 120 is not an outlier.
(III) 121 is an outlier.

(A) Only (I) is true. **(B)** Only (II) is true.
(C) Only (III) is true. **(D)** Only (I) and (III) are true.
(E) (I), (II) and (III) are all true.

(See the solution on page 69.)

11. In measuring the effectiveness of a new drug treatment for cancer patients, 11 patients were tracked after the treatment to see if the cancer returned. The number of years they were cancer free was recorded, and they were given "*N*" for no cancer if they went at least 10 years without a recurrence. The data was: 3.7 9.8 5.5 *N* 1.2 *N* 7.8 8.8 7.5 *N* 2.9. The median of the data is

(A) 6.5 **(B)** 7.8 **(C)** 7.5 **(D)** *N* **(E)** impossible to determine

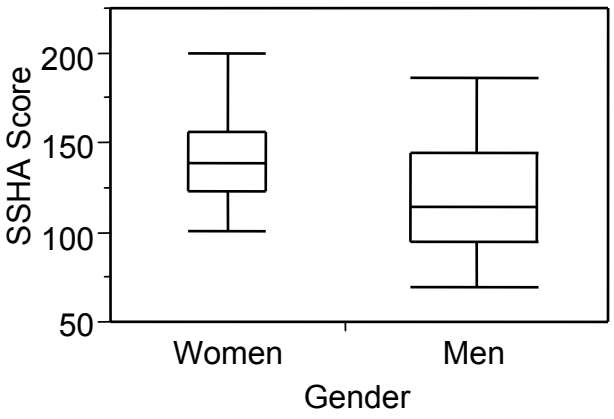
(See the solution on page 70.)

12. The times (in seconds) for 9 subjects to complete a task were as follows: 16 7 3 21 12 7 26 22 45. The interquartile range and variance, respectively, are equal to:

(A) 19; 12.9 **(B)** 17; 12.9 **(C)** 15; 165.5 **(D)** 17; 165.5 **(E)** 19; 165.5

(See the solution on page 72.)

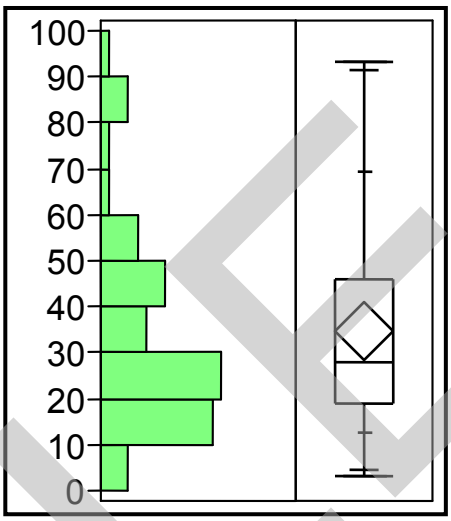
13. The survey of study habits and attitudes (SSHA) is a psychological test that measures the motivation, attitude toward school, and study habits of students. Scores range from 0 to 200. *JMP*™ produced side-by-side boxplots at right to compare the scores on the test for a random sample of men and women at the University of Manitoba. Which statement below is <u>false</u>?



(A) The women's scores have a narrower spread than the men's.
(B) The interquartile range for the men's scores is approximately 50.
(C) More men wrote this test than women.
(D) The women tend to score higher than the men on this test.
(E) The mean test score for the men is less than their median score.

(See the solution on page 73.)

14. At right is the distribution of annual income (in thousands of dollars) of a sample of Canadian adult males as displayed by *JMP*™.



(a) From this information we would conclude:
- **(A)** the sample was clearly not random.
- **(B)** the distribution is symmetrical.
- **(C)** the distribution is left-skewed.
- **(D)** the distribution is right-skewed.
- **(E)** men make more money than women.

(See the solution on page 75.)

(b) The best way to summarize this data is:
- **(A)** the mean and standard deviation.
- **(B)** the mean and variance.
- **(C)** the mean, the median and the mode.
- **(D)** the five-number summary.
- **(E)** none of the above.

(See the solution on page 75.)

(c) The mean and median, respectively, of this distribution are approximately
(A) 28; 35 **(B)** 35; 28 **(C)** 45; 32 **(D)** 32; 45 **(E)** 40; 32

(See the solution on page 76.)

15. The table below shows the scores for the 60 people who participated in a recent 18-hole golf tournament (the winner scored −1 or 1 below par).

Score	−1	0	+1	+2	+3	+4	+5	+6	+7
Frequency	1	0	1	4	3	7	10	25	9

Looking at the distribution of scores, we can conclude:
- **(A)** The median is +3 and the mean is higher.
- **(B)** The median is +3 and the mean is lower.
- **(C)** The median and mean are both +3.
- **(D)** The median is +6 and the mean is lower.
- **(E)** The median and mean are both +6.

(See the solution on page 78.)

16. The stemplot at right shows the number of weeks a sample of 40 patients in Manitoba had to wait before receiving hip replacement surgery. The five-number summary for this data is:

(A) 7, 26, 35, 42, 52
(B) 0, 26.5, 35, 42.5, 69
(C) 7, 26, 35, 42, 69
(D) 7, 26.5, 35, 42.5, 69
(E) 7, 26.5, 35, 40.5, 69

(See the solution on page 79.)

Stem	Leaf
7	
6	9
6	
5	
5	12
4	67789
4	0011234
3	555589
3	13344
2	55667789
2	2
1	589
1	4
0	7

17. The mean daily high temperature for June in a particular city is 20 °C with a standard deviation of 2.7 °C. What would the mean and standard deviation be in Fahrenheit? (Hint: The formula $F = 1.8C + 32$ converts Celsius into Fahrenheit.)

(A) 68 & 3.86 **(B)** 68 & 36.86 **(C)** 68 & 4.86 **(D)** 68 & 34.7
(E) It is impossible to determine without the original data.

(See the solution on page 83.)

18. After a particularly difficult exam, the average mark was 47%, the median was 40%, and the interquartile range was 10%. The prof decides to add 20% to everyone's mark. After this addition, which of the following statements is FALSE?

(A) The first quartile could now be as low as 50%.
(B) The third quartile could now be as high as 70%.
(C) The median mark is now 60%.
(D) The interquartile range is now 30%.
(E) The average mark is now 67%.

(See the solution on page 84.)

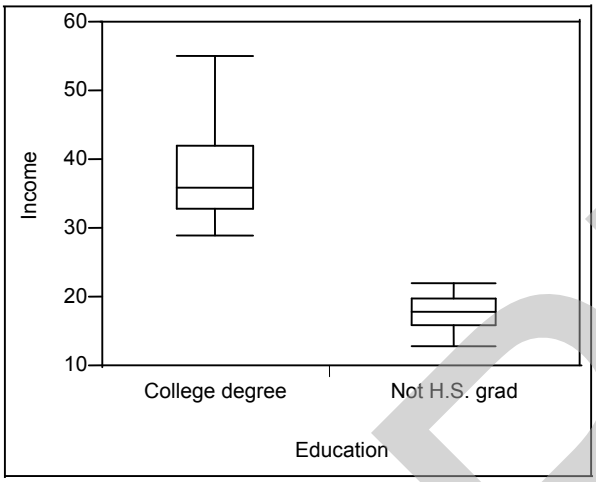
19. A student's average after writing six tests is 68%. She just got her seventh test back with a mark of 82%. What is her average mark now?

(A) 74% **(B)** 71% **(C)** 75% **(D)** 72% **(E)** 70%

(See the solution on page 85.)

20. Below is a *JMP*™ printout comparing the annual income (in thousands of dollars) of two random samples of 35 year-old full-time workers. One sample is of 10 people who have a college degree, the other is 11 people who did not graduate from high school.

Oneway Analysis of Income By Education



Quantiles

Level	Minimum	10%	25%	Median	75%	90%	Maximum
College degree	29	29.1	33	36	42	54.3	55
Not H.S. grad	13	13.2	16	18	20	21.8	22

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
College degree	10	38.1000	7.97844	2.5230	32.393	43.807
Not H.S. grad	11	18.0000	2.82843	0.8528	16.100	19.900

What is the mean and median income of all 21 workers?

(A) \$27,571 and \$27,000, respectively
(B) \$27,571 and \$22,000, respectively
(C) \$27.571 and \$27, respectively
(D) \$28,050 and \$27,000, respectively
(E) \$28,050 and \$22,000, respectively

(See the solution on page 88.)

HOMEWORK FOR LESSON 1

- Study the lesson thoroughly until you can do all of the Lecture Problems from start to finish without any assistance. **I have collected the Lecture Problems together for your convenience starting on page 91 above.**
- I have provided a **Summary of Key Concepts** starting on page 89 above.
- **Do not try to learn the material by doing your hand-in assignments. Learn the lesson first, then use the hand-in assignments to test your understanding of the lesson.** Before each hand-in assignment, I will send you tips telling you what lesson you should be studying to prepare for the assignment. Make sure you sign up for Grant's Homework Help at www.grantstutoring.com to receive these tips.
- If you have the *Multiple-Choice Problems Set for Basic Statistical Analysis I (Stat 1000)* by Smiley Cheng available in the Statistics section of the UM Book Store (and I do recommend you get this book for all the old midterm and final exams contained within it, if nothing else), then additional practise of many of the concepts taught in this lesson is available in:
  - **Section GDS: Graphs and Descriptive Statistics** (**OMIT Question 17**). **The solutions to this section are provided in Appendix B of my book starting on page B-1 below.**
- **Have you signed up for Grant's Homework Help yet?** Tips have already been sent to help you prepare for some of the hand-in assignments. Clear step-by-step instructions on how to get *JMP*™ to do the various things required have been sent as well. It's all FREE! Go to **www.grantstutoring.com** to sign up.

APPENDIX A

HOW TO USE STAT MODES ON YOUR CALCULATOR

In the following pages, I show you how to enter data into your calculator in order to compute the mean and standard deviation. I also show you how to enter x, y data pairs in order to get the correlation, intercept and slope of the least squares regression line.

Please make sure that you are looking at the correct page when learning the steps. I give steps for several brands and models of calculator.

I consider it absolutely vital that a student know how to use the Stat modes on their calculator. It can considerably speed up certain questions and, even if a question insists you show all your work, gives you a quick way to check your answer.

If you cannot find steps for your calculator in this appendix, or cannot get the steps to work for you, do not hesitate to contact me. I am very happy to assist you in calculator usage (or anything else for that matter).

SHARP CALCULATORS
(Note that the EL-510 does not do Linear Regression.)

You will be using a "MODE" button. Look at your calculator. If you have "MODE" actually written on a button, press that when I tell you to press "[MODE]". If you find mode written above a button (some models have mode written above the "DRG" button, like this: "[MODE] [DRG]") then you will have to use the "[2nd F]" button to access the mode button; i.e. when I say "[MODE]" below, you will actually press "[2nd F] [MODE] [DRG]".

BASIC DATA PROBLEM

Feed in data to get the mean, $\bar{x}$, and standard deviation, *s* (which Sharps tend to denote "*sx*").

Step 1: Put yourself into the "STAT, SD" mode.
Press [MODE] [1] [0] (Screen shows "Stat0")

Step 2: Enter the data: 3, 5, 9.
To enter each value, press the "M+" button. There are some newer models of Sharp that have you press the "CHANGE" button instead of the "M+" button. (The "CHANGE" button is found close by the "M+" button.)

3 [M+] (DATA) 5 [M+] (DATA) 9 [M+] (DATA)
You should see the screen counting the data as it is entered (Data Set=1, Data Set=2, Data Set=3).

Step 3: Ask for the mean and standard deviation.
[RCL] [4] ($\bar{x}$)
We see that $\bar{x}$=5.6666...=5.6667.
[RCL] [5] (*sx*)
We see that *s*=3.05505...=3.0551

Step 4: Return to "NORMAL" mode. This clears out your data as well as returning your calculator to normal.
[MODE] [0]

LINEAR REGRESSION PROBLEM

Feed in *x* and *y* data to get the correlation coefficient, r, the intercept, a, and the slope, b.

Step 1: Put yourself into the "STAT, LINE" mode.
Press [MODE] [1] [1] (Screen shows "Stat1")

Step 2: Enter the data:

x	3	5	9
y	7	10	14

Note you are entering in pairs of data (the *x* and *y* must be entered as a pair). The pattern is first *x*, press "STO" to get the comma, first *y,* then press "M+" (or **"CHANGE")** to enter the pair; repeat for each data pair.
3 [STO] (x,y) 7 [M+] (DATA)
5 [STO] (x,y) 10 [M+] (DATA)
9 [STO] (x,y) 14 [M+] (DATA)
You should see the screen counting the data as it is entered (Data Set=1, Data Set=2, Data Set=3).

Step 3: Ask for the correlation coefficient, intercept, and slope. (The symbols may appear above different buttons than I indicate below.)
[RCL] [÷] (r)
We see that r=0.99419...=0.9942.
[RCL] [(] (a)
We see that a=3.85714...=3.8571.
[RCL] [)] (b)
We see that b=1.14285...=1.1429.

Step 4: Return to "NORMAL" mode. This clears out your data as well as returning your calculator to normal.
[MODE] [0]

CASIO CALCULATORS
(Note that some Casios do not do Linear Regression.)

BASIC DATA PROBLEM

Feed in data to get the mean, $\bar{x}$, and standard deviation, *s* (which Casios tend to denote "$x\sigma_{n-1}$" or simply "σ_{n-1}").

Step 1: Put yourself into the "SD" mode.

Press "MODE" once or twice until you see "SD" on the screen menu and then select the number indicated. A little "SD" should then appear on your screen.

Step 2: Clear out old data.

SHIFT AC (Scl) = (Some models will have "Scl" above another button. Be sure you are pressing "Scl", the "Stats Clear" button. (Some models call it "SAC" for "Stats All Clear" instead of Scl.)

Step 3: Enter the data: 3, 5, 9.
To enter each value, press the "M+" button.

3 M+ (DT) 5 M+ (DT) 9 M+ (DT) (You use the "M+" button to enter each piece of data.)

Step 4: Ask for the mean and standard deviation.

SHIFT 1 ($\bar{x}$) =

We see that $\bar{x}$ = 5.6666...=5.6667.

SHIFT 3 ($x\sigma_{n-1}$) =

We see that *s* = 3.05505...=3.0551

(Some models may have $\bar{x}$ and $x\sigma_{n-1}$ above other buttons rather than "1" and "3" as I illustrate above.)

If you can't find these buttons on your calculator, look for a button called "S. VAR" (which stands for "Statistical Variables", it is probably above one of the number buttons).

Press: SHIFT S. VAR and you will be given a menu showing the mean and standard deviation. Select the appropriate number on the menu and press "=" (You may need to use your arrow buttons to locate the $\bar{x}$ or $x\sigma_{n-1}$ .options.)

Step 5: Return to "COMP" mode.

Press MODE and select the "COMP" option.

LINEAR REGRESSION PROBLEM

Feed in *x* and *y* data to get the correlation coefficient, r, the intercept, a, and the slope, b.

Step 1: Put yourself into the "REG, Lin" mode.

Press "MODE" once or twice until you see "Reg" on the screen menu and then select the number indicated. You will then be sent to another menu where you will select "Lin". (Some models call it the "LR" mode in which case you simply choose that instead.)

Step 2: Clear out old data.

Do the same as Step 2 for "Basic Data".

Step 3: Enter the data.

x	3	5	9
y	7	10	14

Note you are entering in pairs of data (the *x* and *y* must be entered as a pair). The pattern is first *x*, first *y*; second *x*, second *y*; and so on. Here is the data we want to enter:

3 , 7 M+ (DT) 5 , 10 M+ (DT) 9 , 14 M+ (DT)

(If you can't find the comma button " , ", you probably use the open bracket button instead to get the comma "[(−−". You might notice "[x_D, y_D]" in blue below this button, confirming that is your comma.)

Step 4: Ask for the correlation coefficient, intercept, and slope. (The symbols may appear above different buttons than I indicate below.)

SHIFT ((r) =

We see that r = 0.99419...=0.9942.

SHIFT 7 (A) =

We see that a = 3.85714...=3.8571.

SHIFT 8 (B) =

We see that b = 1.14285...=1.1429.

If you can't find these buttons on your calculator, look for a button called "S. VAR"

Press: SHIFT S. VAR and you will be given a menu showing the mean and standard deviation. **Use your left and right arrow buttons to see other options**, like "*r*". Select the appropriate number on the menu and press "=".

Step 5: Return to "COMP" mode.

Press MODE and select the "COMP" option.

HEWLETT PACKARD HP 10B II

BASIC DATA PROBLEM

Feed in data to get the mean, $\bar{x}$, and standard deviation, *s* (which it denotes "*Sx*").

Step 1: Enter the data: 3, 5, 9.
To enter each value, press the "Σ+" button.

3 Σ+ 5 Σ+ 9 Σ+ (As you use the "Σ+" button to enter each piece of data, you will see the calculator count it going in: 1, 2, 3.)

Step 2: Ask for the mean and standard deviation.
Note that by "orange" I mean press the button that has the orange bar coloured on it. The orange bar is used to get anything coloured orange on the buttons.

orange 7 ($\bar{x},\bar{y}$)

We see that $\bar{x}$ = 5.6666...=5.6667.

orange 8 (s_x, s_y)

We see that *s* = 3.05505...=3.0551

Step 3: "Clear All" data ready for next time.

orange C (*C ALL*)

LINEAR REGRESSION PROBLEM

Feed in *x* and *y* data to get the correlation coefficient, r, the intercept, a, and the slope, b.

Step 1: Enter the data:

x	3	5	9
y	7	10	14

Note you are entering in pairs of data (the *x* and *y* must be entered as a pair). The pattern is first *x*, first *y*; second *x*, second *y*; and so on.

3 INPUT 7 Σ+

5 INPUT 10 Σ+

9 INPUT 14 Σ+

(As you use the "Σ+" button to enter each pair of data, you will see the calculator count it going in: 1, 2, 3.)

Step 2: Ask for the correlation coefficient, intercept, and slope.

orange 4 ($\hat{x},r$) orange K (*SWAP*)

We see that *r* = 0.99419...=0.9942.
Note that the "SWAP" button is used to get anything that is listed second (after the comma) like "*r*" in this case.

The intercept has to be found by finding $\hat{y}$ when *x*=0:

0 orange 5 ($\hat{y},m$)

We see that *a* = 3.85714...=3.8571.

The slope is denoted "*m*" on this calculator:

orange 5 ($\hat{y},m$) orange K (*SWAP*)

We see that *b* = 1.14285...=1.1429.

Step 3: "Clear All" data ready for next time.

orange C (*C ALL*)

TEXAS INSTRUMENTS TI-30X-II
(Note that the TI-30Xa does not do Linear Regression.)

BASIC DATA PROBLEM

Feed in data to get the mean, $\bar{x}$, and standard deviation, *s* (which it denotes "*Sx*").

Step 1: Clear old data.

2nd STAT DATA Use your arrow keys to ensure "*CLRDATA*" is underlined then press ENTER =

Step 2: Put yourself into the "STAT 1-Var" mode.

2nd STAT DATA Use your arrow keys to ensure "1-Var" is underlined then press ENTER =

Step 3: Enter the data: 3, 5, 9.
(You will enter the first piece of data as "X1", then use the down arrows to enter the second piece of data as "X2", and so on.)

DATA 3 ENTER = (X1 = 3)

⇓ ⇓ 5 ENTER = (X2 = 5)

⇓ ⇓ 9 ENTER = (X3 = 9)

Step 4: Ask for the mean and standard deviation.
Press STATVAR then you can see a list of outputs by merely pressing your left and right arrows to underline the various values.

We see that $\bar{x}$ = 5.6666...=5.6667.

We see that *s* = 3.05505...=3.0551

Step 5: Return to standard mode.
CLEAR This resets your calculator ready for new data next time.

LINEAR REGRESSION PROBLEM

Feed in *x* and *y* data to get the correlation coefficient, r, the intercept, a, and the slope, b.

Step 1: Clear old data (as in **BASIC DATA PROBLEM** at left).

Step 2: Put yourself into the "STAT 2-Var" mode.

2nd STAT DATA Use your arrow keys to ensure "2-Var" is underlined then press ENTER =

Step 3: Enter the data:

x	3	5	9
y	7	10	14

(You will enter the first *x*-value as "X1", then use the down arrow to enter the first *y*-value as "Y1", and so on.)

DATA 3 ENTER = ⇓ 7 ENTER = (X1 = 3, Y1 = 7)

⇓ 5 ENTER = ⇓ 10 ENTER = (X2 = 5, Y2 = 10)

⇓ 9 ENTER = ⇓ 14 ENTER = (X3 = 9, Y3 = 14)

Step 4: Ask for the correlation coefficient, intercept, and slope.
Press STATVAR then you can see a list of outputs by merely pressing your left and right arrows to underline the various values. **Note: Your calculator may have a and b reversed. To get a, you ask for b; to get b you ask for a.** Don't ask me why that is, but if that is the case then realize it will <u>always</u> be the case.

We see that r = 0.99419...=0.9942.

We see that a = 3.85714...=3.8571.

We see that b = 1.14285...=1.1429.

Step 5: Return to standard mode (as in **BASIC DATA PROBLEM** at left).

TEXAS INSTRUMENTS TI-36X
(Note that the TI-30Xa does not do Linear Regression.)

BASIC DATA PROBLEM

Feed in data to get the mean, $\bar{x}$, and standard deviation, *s* (which it denotes "σx_{n-1}").

Step 1: Put yourself into the "STAT 1" mode.

3rd [x ⇄ y] (STAT 1)

Step 2: Enter the data: 3, 5, 9.
To enter each value, press the "Σ+" button.

3 [Σ+] 5 [Σ+] 9 [Σ+] (As you use the "Σ+" button to enter each piece of data, you will see the calculator count it going in: 1, 2, 3.)

Step 3: Ask for the mean and standard deviation.

2nd [$\bar{x}$ / x^2]

We see that $\bar{x}$ = 5.6666...=5.6667.

2nd [σ_{xn-1} / $\sqrt{x}$]

We see that *s* = 3.05505...=3.0551

Step 4: Return to standard mode.

[ON/AC] (Be careful! If you ever press this button during your work you will end up resetting your calculator and losing all of your data. Use the [CE/C] button to clear mistakes without resetting your calculator. I usually press this button a couple of times to make sure it has cleared any mistake completely.)

LINEAR REGRESSION PROBLEM

Feed in *x* and *y* data to get the correlation coefficient, r, the intercept, a, and the slope, b.

Step 1: Put yourself into the "STAT 2" mode.

3rd [Σ+] (STAT 2)

Step 2: Enter the data:

x	3	5	9
y	7	10	14

Note you are entering in pairs of data (the *x* and *y* must be entered as a pair). The pattern is first *x*, first *y*; second *x*, second *y*; and so on.

3 [x ⇄ y] 7 [Σ+]

5 [x ⇄ y] 10 [Σ+]

9 [x ⇄ y] 14 [Σ+]

(As you use the "Σ+" button to enter each pair of data, you will see the calculator count it going in: 1, 2, 3.)

Step 3: Ask for the correlation coefficient, intercept, and slope.

Note that this calculator uses the abbreviations "COR" for correlation, "ITC" for intercept and "SLP" for slope.

3rd [4] (COR)

We see that r = 0.99419...=0.9942.

2nd [4] (ITC)

We see that a = 3.85714...=3.8571.

2nd [5] (SLP)

We see that b = 1.14285...=1.1429.

Step 4: Return to standard mode.

[ON/AC]

TEXAS INSTRUMENTS TI-BA II Plus

Put yourself into the "LIN" mode.

2nd [STAT] 8 If "LIN" appears, great; if not, press 2nd [SET] ENTER repeatedly until "LIN" does show up. Then press 2nd [QUIT] CPT to "quit" this screen.

Note: Once you have set the calculator up in "LIN" mode, it will stay in that mode forever. You can now do either "Basic Data" or "Linear Regression" problems.

BASIC DATA PROBLEM

Feed in data to get the mean, $\bar{x}$, and standard deviation, *s* (which it denotes "S*x*").

Step 1: Clear old data.

2nd [DATA] 7 2nd [CLR Work] CE/C

Step 2: Enter the data: 3, 5, 9.

(You will enter the first piece of data as "X1", then use the down arrows to enter the second piece of data as "X2", and so on. Ignore the "Y1", "Y2", etc.)

DATA 3 ENTER = (X1 = 3)

⇓ ⇓ 5 ENTER = (X2 = 5)

⇓ ⇓ 9 ENTER = (X3 = 9)

Step 3: Ask for the mean and standard deviation.

Press 2nd [STAT] 8 then you can see a list of outputs by merely pressing your up and down arrows to reveal the various values.

We see that $\bar{x}$ = 5.6666...=5.6667.

We see that *s* = 3.05505...=3.0551

Step 4: Return to standard mode.

ON/OFF This resets your calculator ready for new data next time.

LINEAR REGRESSION PROBLEM

Feed in *x* and *y* data to get the correlation coefficient, r, the intercept, a, and the slope, b.

Step 1: Clear old data.

2nd [DATA] 7 2nd [CLR Work] CE/C

Step 2: Enter the data:

x	3	5	9
y	7	10	14

(You will enter the first *x*-value as "X1", then use the down arrow to enter the first *y*-value as "Y1", and so on.)

DATA 3 ENTER = ⇓ 7 ENTER = (X1 = 3, Y1 = 7)

⇓ 5 ENTER = ⇓ 10 ENTER = (X2 = 5, Y2 = 10)

⇓ 9 ENTER = ⇓ 14 ENTER = (X3 = 9, Y3 = 14)

Step 3: Ask for the correlation coefficient, intercept, and slope.

Press 2nd [STAT] 8 then you can see a list of outputs by merely pressing your up and down arrows to reveal the various values.
We see that r = 0.99419...=0.9942.

We see that a = 3.85714...=3.8571.

We see that b = 1.14285...=1.1429.

Step 4: Return to standard mode.

ON/OFF This resets your calculator ready for new data next time.