Graphing Calculator Basics for Elementary Statistics

James M. (Mike) Wilkes
Indian River State College
ASC Chastain Campus

November 19, 2019

Contents

Preface

This document was originally created simply as a guide and tutorial on the use of the TI-83 and TI-84 graphing calculators for statistics calculations. It was not intended to replace the actual *learning* of the material required for the Elementary Statistics course STA2023 at Indian River State College. This expanded version has added some notes that may be helpful in understanding the subject, but should be considered as no more than a supplement to the material of the textbook. It has in fact become so unwieldy by now that I expect the probability of being read by any student, or instructor for that matter, is roughly the same as that of reading the actual textbook.

Chapters and their titles correspond to those of reference [1], being used as of this writing in STA2023 (the Fall semester 2014). The material typically covered in the course from chapters 1 and 4 does not require a graphing calculator. Our presentation of calculator applications in the remaining course chapters 2–9, and 13, is by no means comprehensive. For example, included under Chapter 5 on discrete distributions are calculations involving only the binomial distribution, and in Chapter 6 on continuous distributions only those for the normal distribution; we have been similarly selective of topics from the other chapters, as will be obvious. It came to our attention after completing this work that a detailed manual on graphing calculator applications written to accompany the older 6th edition of Mann's textbook can be read online, or downloaded, from the website `http://spaces.imperial.edu/rick.castrapel/files/m119/ti83_84_manualMann.pdf` It is more than three times the length of this document, including a number of topics not covered in STA2023, but an excellent reference for details and additional applications.

For Chapter 9 we have attempted to summarize some of the main points concerning hypothesis testing before discussing calculator applications. Impatient readers may wish to skip to the last subsection of that chapter to begin using their calculators, but output from the calculator applications does *not* include the decision to reject or accept an hypothesis. That final step requires understanding of material either from the textbook, or from our summary preceding the calculator applications to effectively *use* the calculator results in making this decision.

There are sometimes subtle differences between operations on the TI-83 and TI-84 calculators, which we take some care to point out; there is at least one useful application on the TI-84 that is *not* on the TI-83 (the inverse t distribution, **invT**), and directions are given for obtaining such a program for the TI-83.

Chapter 1: Introduction

A graphing calculator is not required for the first chapter. The chapter introduces many terms and their definitions, however, that will be used throughout the course. You should read carefully the definitions of *descriptive* and *inferential* statistics, and understand how they are different. Basically, descriptive statistics involves methods of organizing and displaying data using tables and graphs, and calculating measures of the data that attempt to characterize a sometimes immense number of data points by just a few numbers, such as the mean, median, mode, range, variance, and standard deviation (which are discussed in detail in chapters that follow). Inferential statistics on the other hand introduces the ideas of *population* data, and subsets (smaller sets selected from the population) of this data called *sample* data. The goal of inferential statistics is to attempt to derive information about the the population from the smaller sample set, and this requires the development of methods that allow us to *infer* such population information using the sample data. The subjects of probability and probability distributions provide a link between descriptive and inferential statistics, and are discussed in Chapters 4 through 7. Chapter 8 and the chapters following it delve into some of the methods used in inferential statistics.

One problem that seems to occur often, especially in the online WileyPlus homework problems, is that it is difficult to discern whether a given data set represents a population, or just a sample from some population. At this writing (7/22/15) the author does not have a foolproof answer to this problem. Generally, if the word “all” is used in connection with the data set, then it likely represents a population, and if you see the word “sample” in connection with the data set, then it represents a sample from a population. If these words do not appear in the problem statement, then flip a coin to make your decision between population and sample (sorry, that is the best I can offer at the moment).

Chapter 2: Organizing and Graphing Data

- Population or sample data are entered into a list by pressing the [STAT] key, then under the [EDIT] tab selecting [1:Edit]. This places you at the first element of list L1, where you can begin entering data (see next paragraph). To go to another list, scroll *right* using the [▶] key to one of the five other lists L2, L3, L4, L5, or L6 (NOTE: scroll *left* by using the [◀] key, scroll *up* using the [▲] key, or scroll *down* using the [▼] key).

  You may find that you need to *clear* previously entered data from a list before entering new data: to do this, scroll up to the *name* of the given list, for example, L1, L2, etc., press the [CLEAR] key, then press the [ENTER] key at the bottom right corner of the keyboard. This returns you to the first entry of the list, and you should find that all the previous data have disappeared.

  To enter new data in an empty list, go to the first position under the list name. In the window *below* the list will appear, for example, L1(1)=, if you are in list L1. Use the keyboard to input the number you want and press [ENTER]. The number will be placed in the first entry of the list, and cursor repositioned at the second entry point. Repeat until all of your data have been entered. Press [2ND], then [MODE], to exit from [STAT] and return to the *homescreen* (the main window where calculations are usually performed). To return to the homescreen from any application, you can always press [2ND] then [MODE] to quit the application (notice the word QUIT above the [MODE] key).

  Some applications/programs will require the *input of a list*. Pressing the [2ND] key, then the number [1] key, will input the list L1 (L1 appears above the [1] key) *at the point of insertion of input*, and similarly for the other number keys [2] thru [6] to input lists L2 thru L6 on the homescreen.

- *Raw* (ungrouped) *data* require *one* list, say, L1.

- *Class* (grouped) *data* require *two* lists, say, L1 and L2, the first containing the class *midpoints*, the second containing the corresponding class *frequencies*. You may have to calculate the midpoints as "midpoint = (upper limit + lower limit)/2", if they are not given in the problem statement. Read the textbook *carefully* . . . there is a subtle, but important, difference between the definitions of upper and lower *limits*, and upper and lower *boundaries* (the upper boundary of one class is the *same as* the lower boundary of the *next* class, not so with the upper and lower *limits*). To remember this, *boundaries* are continuous, but *limits* have gaps. See Section 2.2.1 of Reference [1].

Chapter 3: Numerical Descriptive Measures

Given either ungrouped or grouped data for a single variable, we typically want to compute (i) measures of central tendency like the *mean*, the *median* , and the *mode*, (ii) measures of the dispersion or spread in the data, like the *variance* and *standard deviation* (which is the square root of the variance), and (iii) the first and third quartiles Q_1 and Q_3 (used to calculate the interquartile range $IQR = Q_3 - Q_1$).

All of these statistics *except* the mode and variance can be found by first entering your *ungrouped* data into L1, or for grouped data entering the *midpoints* into L1 and *frequencies* into L2. After entering the data:

1. For either the TI-83 or TI-84, press the [STAT] key, scroll over to the [CALC] column, scroll to **1-Var Stats** and press [ENTER].

2. On the TI-84, for ungrouped data in L1, input L1 for `List` and leave `FreqList` empty. For grouped data with midpoints in L1 and frequencies in L2, input L1 for `List`, and L2 for `FreqList`.

3. On the TI-83, the function **1-Var Stats** will appear on your homescreen, and you must then input either L1, or the comma–separated lists L1,L2 for grouped data.

4. Press [ENTER]. The *output* will include the mean $\bar{\texttt{x}}$, the *sample* standard deviation `Sx`, the *population* standard deviation σ`x`, the sample size `n`, the minimum and maximum values, `minX` and `maxX`, of the data, the first and third quartiles `Q1` and `Q3`, and the median, `Med` (which is also the second quartile, Q_2). Both sample and population standard deviations are displayed, since the calculator has no way of knowing whether the data is from a population, or a sample of a population. Only *you* can determine *which type* of data you have in order to choose the correct standard deviation.

5. The sample and population variances are *not* displayed, but can be obtained from the `Sx` and σ`x` standard deviations output from **1-Var Stats** by pressing [VARS], selecting `Statistics` from the VARS tab, then selecting either `Sx` or σ`x` from the `XY` tab; `Sx` or σ`x` will appear on the homescreen, and can be squared to get the sample or population variance by pressing the [x^2] key.

6. To compute the sample variance directly, or if you need only a *single* statistic of the data, like the mean or standard deviation: press [2ND], then [STAT] (putting you in the LIST menu, as indicated above the [STAT] key), scroll right to the `MATH` tab, then find the *statistic* you want in the list that is displayed. For example, if you wish to compute the *sample variance* of your data (be warned, there is *no option* under the `MATH` tab to calculate the *population variance*), select **variance**. This places **variance** on your homescreen, where you must then input the list of data, say, L1, by pressing [2ND], then the number [1] key. For *grouped* data, input the comma–separated lists L1,L2. The output is the variance of your data.

Your calculator does not provide an application to determine *directly* the *mode* or *modes* of a data set. However, you *can* rank-order your list in either ascending (smallest to largest) order, or descending (largest to smallest) order, by pressing [2ND], [STAT] (putting you in the LIST menu), then scrolling to the `OPS` tab. If you select **SortA**, then press [2ND] followed by [1] to input L1 in the command that appears on the homescreen, it will *replace* your list L1 with the *same* data sorted from smallest to largest. For grouped data, you should input the frequency list L2 instead, which replaces L2 by the same frequency data sorted from smallest to largest. The *mode* of the data is either the data element that appears the *most* number of times in your rank-ordered data list L1, or the midpoint with *highest frequency* that appears in your rank-ordered frequency list L2 (if more than one data element appears the *same* maximum number of times, or more than one midpoint appears with a frequency having the *same* maximum value, then there will be more than one mode). Note that by choosing **SortD** from the `OPS` tab, you obtain data ordered from largest to smallest.

Programs *have* been written that will calculate *directly* the mode, or modes, of a list of raw data. One of these can be downloaded from the Calcblog website:

`http://www.calcblog.com/resources`

by scrolling down to Downloads, looking under Calcblog Software, and selecting the *TI-83 and TI-84 Mode Program* link. This program is written in TI-calculator *assembly language*, but it can be saved to your computer as a text file (that will be basically incomprehensible), and then transferred to your calculator using the TI Connect software. Instructions for making the transfer can be found on the Calcblog website, above, or on the TI website

`http://www.ticalc.org`

under the ”Help” link, where you should select the ”Learn the Basics” link. Another link to detailed instructions can be found at

`http://mikewilkes-irsc.weebly.com/probability-and-statistics.html`

where the file is also available for download.

Mean, Standard Deviation, and Variance on Non-Graphing TI Calculators

This topic lies outside the primary objective of discussing only TI graphing calculator applications, but may be of interest to students who would like to analyze data on older, non-graphing Texas Instrument calculators.

- TI-30X. First press the [2nd] key, then the [7] key (with the entry CSR above it, standing for Clear Statistics Register) to *clear* any previously entered data (if no data has been entered you will get an ERROR message . . . just ignore this and go on). For raw data, key in a data value; for class data, key in the midpoint of a class. If the data value occurs with frequency greater than 1, or if it is a class midpoint, press [2nd] then [$1/x$] (FRQ is above this key) and enter the frequency of this data point or class. For data points of frequency 1, that is, single distinct values of raw data, this step can be ignored. Finally, press the [$\Sigma+$] key to complete entry of the data point/class midpoint, and its frequency, into the calculator. The calculator will echo $n =$ the number of data values entered so far (equal to the *sum* of the frequencies entered). To remove an incorrectly entered data value and frequency, press the [$\Sigma-$] key. Continue entering one point at a time, pressing [2nd] then [$1/x$] to enter the associated frequency, then pressing the [$\Sigma+$] key to enter the point and its frequency into the calculator. After entering all the data points/class midpoints, press [2nd] then the [x^2] key to find the mean value $\overline{x}$ (which appears above the [x^2] key), press [2nd] then the [$\sqrt{x}$] key to find the sample standard deviation (σ_{xn-1} is above this key), or press [2nd] then the division key [$\div$] to find the population standard deviation (σ_{xn} is above this key). To find the sample variance, press [2nd] then [$\sqrt{x}$] to first obtain the sample standard deviation, then press the [x^2] key to find the sample variance (it is, by definition, the square of the sample standard deviation). Similar steps will give you the population variance.

- TI-35X. Data entry and calculations are very similar to those for the TI-30X, except we *begin* by pressing the [3rd] key, then the [$x\rightleftarrows y$] key (above which you will find STAT 1), to enter single variable data entry mode. Then press [2nd] followed by the [$x\rightleftarrows y$] key to clear any previous data (notice that CSR appears above $x\rightleftarrows y$ on this key). Ignore any ERROR message, it just means there was no previous data entered. Begin entering data points exactly as was done for the TI-30X, that is, enter a data point/class midpoint, then press [2nd] followed by the [$1/x$] key (which shows FRQ above it) to enter the associated frequency. Press the [$\Sigma+$] key to complete entry of the data point/class midpoint, and its frequency, into the calculator. Repeat until all data points/class midpoints, and their frequencies, have been entered. Then, just as for the TI-30X, press [2nd] then the [x^2] key to find the mean value $\overline{x}$, press [2nd] then the [$\sqrt{x}$] key to find the sample standard deviation, or press [2nd] then the division key [$\div$] to find the population standard deviation. To find the population variance, press [2nd] then [$\div$] to first obtain the population standard deviation, then press the [x^2] key to find the population variance (it is, by definition, the square of the population standard deviation). Similar steps will give you the sample variance, as detailed above for the TI-30X.

- TI-30X IIS. Both raw and class data can be input to determine the mean, standard deviation, and variance of the data. Begin by pressing [2nd], then [DATA] (the entry above this key is [STAT]), scroll right to CLRDATA, then press [ENTER] to clear any previously entered data. You will be returned to a highlighted 1-VAR, in which case press [ENTER] again, to enter single variable data entry mode. Now press [DATA] to begin entering data at X1=. If it is raw data, enter the data point, scroll down to FRQ=, either leave this empty or enter the number of times this point occurs in the data set, then scroll down again to X2=, and enter the next data point. Continue until all data points have been entered. If it is class data, enter the *midpoint* of class 1 at the X1= prompt, scroll down to FRQ=, and enter the frequency associated with this class, then scroll down once more to X2=, and enter the second class midpoint. Scroll down to enter at FRQ= the frequency associated with the second class of your data. Continue until all class midpoint and frequency data have been entered. As the final step, press the [STATVAR] key which displays on the first line: n, the number of data points, or the sum of all frequencies; $\overline{x}$, the mean of the data; Sx, the standard deviation of a *sample* survey data set; σx, the population standard deviation of a census data set; Σx, the sum of all the x-values; and, Σx^2, the sum of the squared data values. Scroll to any one of these outputs, and its value will be displayed on the second line of the screen. To calculate the *sample variance*, scroll to the right to Sx, then press the [x^2] key followed by [ENTER] to display its value. For the *population* variance, scroll instead to σx, then press the [x^2] key followed by [ENTER] to display its value.

Chapter 4: Probability

A discussion of problems involving probability requires a large number of definitions, which we simply itemize and discuss very briefly here.

Experiments, Observations, and Outcomes

Experiment – a procedure or act of making an *observation*, which can be repeated under precisely the same conditions. Simple experiments that are discussed often in the text include flipping a coin, either once or a repeated number of times, and rolling a six-sided die (a "die" is one of a pair of dice), either one or more times.

Outcome – the result of an observation. An outcome of flipping a coin once is either a "head" or a "tail", symbolized by an H or T. An outcome of flipping a coin twice could be one of the following four pairs, either HH, HT, TH, or TT. An outcome of tossing a die is one of the six numbers $1, 2, 3, 4, 5$, or 6.

Random Experiment – an experiment whose outcomes cannot be predicted with certainty prior to the performance of the experiment. However, the collection or set of all *possible* outcomes *is assumed to be known* prior to the performance of the experiment. It is further assumed that at least in principle the experiment can be repeated indefinitely under the same conditions, and that outcomes of a repeat of an experiment are unaffected by previous performances of the experiment.

Sample Space (of an Experiment) – a set, denoted by S, containing *all possible outcomes* of an experiment. The sample space for the flip of a coin is the set of two possible outcomes: $S = \{H, T\}$. The sample space for flipping a coin twice in succession is $S = \{HH, HT, TH, TT\}$. The sample space for a single roll of a die is $S = \{1, 2, 3, 4, 5, 6\}$.

Countable Sample Space – a sample space S containing a countable number of outcomes: $S = \{o_1, o_2, o_3, \ldots,\}$, where o_i is the ith outcome (the sequence of three "dots" terminating with a comma, called an *ellipsis*, indicates that there is no upper bound on the number of outcomes). For a finite, countable sample space with n outcomes: $S = \{o_1, o_2, o_3, \ldots, o_n\}$.

Event – any *subset* of S (including the *improper* subset S itself); that is, any *set* of one or more *outcomes* of an experiment. The event for which there is *no outcome* is called the *impossible* event, since no outcome is possible. It is represented by the *empty set* $\emptyset$, since it contains no outcomes. The sample space S is also an event, containing every possible event, and is called the *certain* or *sure* event since *some* outcome is certain to occur.

Simple Event – an event containing only *one* outcome of an experiment. If the sample space is countable, the ith simple event is usually denoted $E_i = \{o_i\},\ i = 1, 2, 3, \ldots,$. If the sample space is finite and countable, with n possible outcomes, then $E_i = \{o_i\},\ i = 1, 2, 3, \ldots, n$

Compound Event – an event containing *more than one* outcome of an experiment. For example, in a countable sample space, a possible event containing three outcomes would be $A = \{o_2,\, o_4,\, o_9\}$.

Algebra of Events – since events are *sets*, the set operations of union and intersection apply to them:

Union of Events – for any events A and B, their union $A \cup B$ is the set of all outcomes that are in A, together with all those in B not already included in A. Stated differently, $A \cup B$ is the set of outcomes belonging to either A or B, or both A and B (the event is the empty set $\emptyset$ if there are no outcomes in either set A or B). Note that any outcome *common* to both A and B appears only *once* in $A \cup B$ (it should not be counted twice).

The union $A \cup B$ is often read as the logical disjunction "($A\,\text{or}\,B$)", where A is a statement defining event A, and B is a statement defining B. Recall from truth tables that ($A\,\text{or}\,B$) is true when either A or B, or both, are true (and false only when both are false), so we have the natural correspondence

$$A \cup B \;=\; A\,\text{or}\,B.$$

Intersection of Events – for any events A and B, their intersection $A \cap B$ is the set of *only* those outcomes *common to both* A and B.

The intersection $A \cap B$ is often read as the logical conjunction "($A\,\text{and}\,B$)", where A and B are now understood to be statements defining the events A and B. Recall from truth tables that ($A\,\text{and}\,B$) is true only when both A and B are true, and false otherwise, so we have the natural correspondence

$$A \cap B \;=\; A\,\text{and}\,B.$$

Complement of an Event – the event $\overline{A}$ containing only those events of the sample space S that do *not* occur in event A. The complement $\overline{A}$ of A in S is sometimes read "*not* A", and if A is a statement defining event A, then its complement is the logical negation $\sim A$, which is false if A is true, and true if A is false. The correspondence between set complement and logical negation of statements is thus

$$\overline{A} \;= \sim A.$$

Union of an Event and Its Complement – the union of event A and its complement $\overline{A}$ is the entire sample space S, since every outcome must be either in A or not in A (in which case it is in its complement $\overline{A}$). Therefore, $S \;=\; A \cup \overline{A}$. This is equivalent to the logical statement that ($A\,\text{or} \sim A$) is always true (it is always true that a statement is either true or it is false).

Mutually Exclusive Events – two events A and B are *mutually exclusive*, if they have *no events in common*, that is, if their intersection is empty: $A \cap B = \emptyset$. This is the equivalent to the logical statement that ($A\,\text{and} \sim A$) is always false (it is always false that a statement is both true and false).

Simple Events of Countable Sample Spaces – simple events E_i of a countable sample space are mutually exclusive, $E_i \cap E_j = \emptyset$ for all values of i and j, since every outcome appears in one and only one simple event. The union of simple events of a countable sample space is equal to the sample space S itself: $S = E_1 \cup E_2 \cup E_3 \ldots$. For a *finite* countable sample space of n outcomes, S is the union of the finite number of simple events: $S = E_1 \cup E_2 \cup E_3 \ldots \cup E_n$.

Elements of Probability Theory

Again, we list here only the fundamental definitions and properties required for the applications to problems involving probability in the textbook.

Probability– a numerical measure, denoted $P(A)$, of the likelihood that a specific event A will occur.

Conditional Probability – it is understood in writing $P(A)$ that the sample space, that is, the certain event S, is known or given, and it *occurs*, that is, it is assumed that at least one true statement S is know about the experiment under consideration. To *convey this specific information*, we might instead write the probability of A given that S has occurred, or given that statement S is true, as $P(A|S)$, and *call* this the *conditional probability* of event A, given S. This more informative way of stating probabilities based on the sample space does *not*, however, appear in either your textbook [2] or in few, if any, other mainstream textbooks.

On the other hand, if information *in addition to* that given by the sample space S is obtained, then the probabilities of all events must change to reflect the new information. For example, if it is known that a *particular* event B of the sample space has in fact occurred, or equivalently that a particular statement B about the experiment is certainly true, then since S is certain to occur it follows that B *and* S occur, that is, the event $B \cap S = (B \, \mathrm{and} \, S)$ occurs. We must then calculate the probability of every event under the condition that $B \cap S$ has occurred, which effectively makes the event $B \cap S$ the *new sample space* to use in calculating probabilities. But, since B is a subset of S, it is in fact true that $B \cap S = B$, the outcomes of $B \cap S$ are simply those of B. Thus, if it is known that B has occurred, the probability of any other event A is written as $P(A|B)$, since $P(A|B \cap S) = P(A|B)$. This is is read "the conditional probability of event A, given that event B has occurred".

Basic Rules for any Measure of Probability – A proposed measure of probability must always satisfy the following rules:

1. The probability of any event A is *never negative*: $P(A) \geq 0$ for any event A.
2. The probability of the *certain* event S is 1: $P(S) = 1$. Since all outcomes contained in the sample space S have some degree of uncertainty, the only certain event of an experiment is the entire sample space S.

3. For any event A, with complement $\overline{A}$, we must have

$$P(A) + P(\overline{A}) \;=\; 1\,.$$

4. The probability $P(A\cap B)$ that both A and B occur, called the *joint probability* of A and B and also written as $P(A \; and \; B)$, is equal to the *product* of the conditional probability of A given B, and the probability of B:

$$P(A\cap B) \;=\; P(A|B)\,P(B)\,, \quad \text{or} \quad P(A \; and \; B) \;=\; P(A|B)\,P(B)\,. \tag{1}$$

Important Results of the Probability Rules – For any probability measure satisfying the above rules, it can be shown that:

- The probability of the impossible event (represented by the empty set $\emptyset$) is zero:

$$P(\emptyset) = 0\,. \tag{2}$$

- The probability $P(A\cup B) \;=\; P(A \text{ or } B)$ that either A or B, or both events occur, is given by

$$P(A \text{ or } B) \;=\; P(A) + P(B) - P(A \text{ and } B)\,. \tag{3}$$

- For a finite, countable sample space of n outcomes, it follows from Rule 2 and the two previous results that, since $S = E_1 \cup E_2 \cup E_3 \ldots \cup E_n$, and the E_i are mutually exclusive ($E_i \cap E_j = \emptyset$ for any i and j), the sum of the probabilities of the simple events E_1, E_2, $E_3\,, \ldots E_n$ is always 1:

$$P(S) \;=\; P(E_1) + P(E_2) + P(E_3) + \ldots + P(E_n) \;=\; \sum_{i=1}^{n} P(E_i) \;=\; 1\,. \tag{4}$$

Assignment of Probabilities to Events – the crucial step in proceeding with the calculation of probabilities is choosing a method for *assigning* a probability to an event. The three common methods are:

1. **Classical, or Theoretical**: All simple events are assumed to be equally probably, so this could be called a theory of equiprobable events. This assumption, together with equation (4), allows us to determine the probability of a simple event for any countable sample space. For example, in flipping a coin, we nearly always assume that the probability of obtaining a head is the same as that of obtaining a tail: $P(H) \;=\; P(T)$, where now H and T are shorthand for the simple events $\{H\}$ and $\{T\}$. In this case, according to equation (4), $P(H) + P(T) = P(H) + P(H) = P(T) + P(T) = 1$, it follows that $P(H) = P(T) = 1/2$. If there are n simple events in the sample space, the classical probability of a simple event is just $\dfrac{1}{n}$.

2. **Relative Frequency, or Empirical**: The relative frequency approach assigns probabilities empirically based on actual data, the relative frequency of an outcome in a series of identical random experiments is taken to be the probability of the outcome. It is more widely applicable than the Classical approach, since it doesn't require the sample space to consist of equally likely simple events.
3. **Subjective**: In my opinion, the textbook [2] and nearly all other mainstream textbooks on probability and statistics do not do this concept justice. If you are interested in understanding it you should read the work of masters of the approach, like Cox [3], Jaynes [4, 5], and Jeffreys [6] (but especially any of the work of E. T. Jaynes). It is prominent in the literature under the heading Bayesian Statistics.
4. **When is a Relative Frequency Probability Exact?**

   If, say, 450 of *all* 500 customers of a business purchased one or more of a particular item, then the relative frequency $450/500 = 0.9$ is the theoretical or exact probability that a randomly selected customer purchased an item. This is because this is an exact proportion of a *population*.

   On the other hand, if a survey was done of all customers, and 450 of 500 *surveyed* customers purchased an item, then $450/500 = 0.9$ would be an estimate or approximate probability based on an approximate proportion from a *sample* of customers, rather than *all* customers (we would not know the exact proportion, since there may be more customers than were surveyed).

Two-Way Classification Tables

In two-way classification tables, there are always two pairs of *mutually exclusive events*, say M and F, and Y and N. In each pair, either one or the other occurs, but never both. However, each event from one pair can occur simultaneously with either event of the *other* pair. For example, both M and Y can occur, and we let $n(M \,\text{and}\, Y) = n(Y \,\text{and}\, M)$ be the *number of occurrences* of both event M and event Y, that is, their *joint number* of occurrences. Of course, since M and F, and Y and N, are mutually exclusive, the number of joint occurrences of either of these pairs is always zero: $n(M \,\text{and}\, F) = n(F \,\text{and}\, M) = 0$, and $n(Y \,\text{and}\, N) = n(N \,\text{and}\, Y) = 0$. In the tables below, we abbreviate, for example $n(M \,\text{and}\, Y) = n(Y \,\text{and}\, M)$ as simply $n(MY) = n(YM)$, and so on. So, a two-way classification table is really just a tabulation of the number of occurrences of *four possible joint events* from two pairs of mutually exclusive events. It will typically be given to you in the generic form

	M	F
Y	$n(MY)$	$n(FY)$
N	$n(MN)$	$n(FN)$

You will want to first fill in the *totals* of the rows and columns, as shown in the completed Table 1, below:

Table 1: Generic Two-Way Classification Table

Event	M	F	Totals
Y	$n(MY)$	$n(FY)$	$n(MY)+n(FY)=n(Y)$
N	$n(MN)$	$n(FN)$	$n(MN)+n(FN)=n(N)$
Totals	$n(MY)+n(MN)$ $=n(M)$	$n(FY)+n(FN)$ $=n(F)$	$n(MY)+n(FY)+n(MN)+n(FN)$ $=\ n(M)+n(F)\ =\ n(Y)+n(N)$

From the table, we see that in the row and column totals, the number of occurrences of events M, F, Y, and N are given by the sums

$$\begin{aligned} n(M) &= n(MY)+n(MN), && (5)\\ n(F) &= n(FY)+n(FN), && (6)\\ n(Y) &= n(MY)+n(FY), && (7)\\ n(N) &= n(MN)+n(FN), && (8) \end{aligned}$$

and the *total* number N of occurrences of all four events can be calculated in any one of three ways:

$$\begin{aligned} N &= n(MY)+n(FY)+n(MN)+n(FN) && (9)\\ &= n(M)+n(F) && (10)\\ &= n(Y)+n(N). && (11) \end{aligned}$$

The probabilities of unconditional events are based on this total number of occurrences, and calculated as *classical* probabilities. For example,

$$P(M)=\frac{n(M)}{N}, \quad P(Y)=\frac{n(Y)}{N}$$

are the probabilities of events M and Y, and

$$P(MY)=\frac{n(MY)}{N}, \quad P(FN)=\frac{n(FN)}{N}$$

are the *joint* probabilities of $(M \,\text{and}\, Y)$, and $(F \,\text{and}\, N)$, respectively. Notice that the numbers *given* in the original two-way classification are used to calculate *joint* probabilities.

If you are asked to calculate a *conditional* probability, like $P(M|Y)$, we use the formula $P(M|Y)=P(M \,\text{and}\, Y)/P(Y)=$.

$$P(M|Y)=\frac{\dfrac{n(MY)}{N}}{\dfrac{n(Y)}{N}}=\frac{n(MY)}{n(Y)}.$$

Chapter 5: Discrete Random Variables and Their Probability Distributions

A *discrete* random variable is a real-valued function defined on the events of a sample space that can take on only a *finite* number, or *countably infinite* number, of possible values x. See Chapter 4 of the textbook for definitions of the sample space of outcomes and events. For calculator applications, we will be concerned only with a *finite* number of values of x. Each value is assigned a probability $P(x)$, and the set of all such probabilities defines the probability *distribution* of x. These probabilities may either be given, or calculated as *relative frequencies* defined by the population size and given *frequencies* of x. In a typical problem, the x values will be listed in a column (or row), and the corresponding probabilities $P(x)$ in a second column (or row).

The mean μ (this is a letter of the Greek alphabet, pronounced ”mu”) of a random variable x is defined by

$$\mu = \sum_x x\,P(x), \tag{12}$$

and its standard deviation σ (the Greek letter pronounced ”sigma”) is defined by

$$\sigma = \sqrt{\left[\sum_x x^2\,P(x)\right] - \mu^2}\,. \tag{13}$$

Similarly to Chapter 3, these parameters can be calculated using either the TI-83 or TI-84 by creating a list L1 containing the values of x, and a list L2 containing the associated *probabilities* (or relative frequencies) $P(x)$. Then press the [STAT] key, scroll to the [CALC] tab, select **1-Var Stats** and press [ENTER]. On the TI-84, input list L1 for `List` and list L2 for `FreqList` or, on the TI-83, input the comma-separated lists L1, L2. Press [ENTER], and the mean μ of the random variable will be the number output as $\bar{\texttt{x}}$, while the standard deviation will be the number output as σx. Note that there is no output for the sample standard deviation `Sx`, as the probabilities are associated with the random variable of a *population* probability distribution.

We discuss in what follows only the important discrete probability distribution known as the *binomial distribution*. It applies to an experiment involving n identical and independent trials, with only two outcomes possible at each trial. The probabilities of the two outcomes are p and $q = 1 - p$. The probability p must be *given*, and is associated with some *desired* outcome of the two possible outcomes. The binomial probability distribution function gives the probability $P(x)$ of obtaining x desired outcomes out of n trials:

$$P(x) = {}_nC_x\,p^x\,q^{n-x}, \quad \text{where} \quad q = 1 - p. \tag{14}$$

Table I of Appendix C tabulates these probabilities for selected values of n, x, and p.

1. To compute such a probability using the TI-84 calculator, press [2ND], then [VARS] (where the `DISTR` menu appears). Scroll down the `DISTR` menu to **binompdf** and press [ENTER]. In the window that appears, input the number n for `trials`, the

desired-outcome probability p for `p`, and the number of desired outcomes x for `x value`, then scroll down to highlight **paste**, and press [ENTER]. This will bring you to the homescreen with everything filled in to **binompdf**. Press [ENTER] again to output the probability defined by equation (14).

2. For the TI-83, after selecting **binompdf** from `DISTR` and pressing [ENTER], **binompdf** will appear on your homescreen and you will have to input the comma–separated list containing n, p and x, then press [ENTER] to output the probability defined by (14). It will appear on the homescreen as simply **binompdf(** n, p, x **)**, standing for
$$P(x) \;=\; \textbf{binompdf(}\, n,\, p,\, x\, \textbf{)}.$$

3. The mean μ, and standard deviation σ, *of a binomial distribution only* are given by the simple formulas:
$$\mu \;=\; np, \qquad \text{and} \qquad \sigma = \sqrt{npq}\,, \quad \text{where} \quad q \;=\; 1 \;-\; p.$$
These are easily calculated on the homescreen using the multiplication operator, together with the square root function (which you get by the keystrokes [2ND], then [x^2] …notice the $\sqrt{\ }$ symbol above this key). So to calculate σ when, for example, $n = 100$ and $p = 0.3$, press [2ND], then [x^2], then complete the homescreen output with $\sqrt{(100 * 0.3 * (1 - 0.3))}$, and press [ENTER]. Be *careful* about inserting parentheses, there should always be the same number of right parentheses as left parentheses!

4. Some of the textbook and computer exercises for the binomial distribution involve answering questions such as "what is the probability that *at most* x of the n outcomes occur?", or "what is the probability that *at least* x of the n outcomes occur?", or "what is the probability that x to y of the n outcomes occur?". These calculations involve the *cumulative* binomial distribution function, a summation over some *range* of the probabilities $P(x)$. The three types of questions are stated as cumulative probabilities as follows (where X is used for the summation variable):
$$P(\text{at most } x) \;=\; P(X \le x) \;=\; \sum_{X=0}^{x} P(X),$$
$$P(\text{at least } x) \;=\; P(X \ge x) \;=\; \sum_{X=x}^{n} P(X) \;=\; 1 - P(X \le x-1) \;=\; 1 - \sum_{X=0}^{x-1} P(X),$$
$$P(x \text{ to } y) \;=\; P(x \le X \le y) \;=\; \sum_{X=x}^{y} P(X) \;=\; P(X \le y) \;-\; P(X \le x-1).$$
Both the TI-83 and TI-84 include the application **binomcdf** (notice the "c" in the name, standing for "cumulative") that will compute $P(\text{at most } x) \;=\; P(X \le x)$, which can thus be used to answer any of the above three questions. To find the application, press [2ND], then [VARS]. Scroll down the `DISTR` menu to **binomcdf**

and press [ENTER]. On the TI-84 input n for `trials`, p for `p`, and either x, y, or $(x-1)$ for `x value`, depending on which of the three problems you are trying to solve. For the TI-83, you must input a comma–separated list n, p, x, [or $(x-1)$, or y, depending on the question] as shown below:

$$P(\text{at most } x) \;=\; \textbf{binomcdf}(n, p, x),$$

$$P(\text{at least } x) \;=\; 1 \;-\; \textbf{binomcdf}(n, p, x-1),$$

$$P(x \text{ to } y) \;=\; \textbf{binomcdf}(n, p, y) \;-\; \textbf{binomcdf}(n, p, x-1).$$

Notice that in the last two formulas, you must input $(x-1)$ as the *third* argument to **binomcdf**.

The right-hand sides of each of these formulas must be entered in *a single line entry* on the homescreen. For example, in the third formula, you will obtain an *incorrect result* if you select **binomcdf**(n, p, y) and press [ENTER], then on the next line use the subtraction key [−] and select **binomcdf**$(n, p, x-1)$ ... in which case the result will appear on the homescreen as `ANS` − **binomcdf**$(n, p, x-1)$ ... and then press [ENTER]. The result of this order of operations is just the output from the *last* command, which was **binomcdf**$(n, p, x-1)$, definitely not the desired result. Instead, select **binomcdf**(n, p, y), press [−], select **binomcdf**$(n, p, x-1)$, and *only then* press the [ENTER] key.

As an example of the use of the cumulative binomial distribution function, we solve Problem 5.51 of the textbook (using technology rather than Table I of Appendix C):

5.51 According to a Wakefield Research survey of adult women, 50% of the women said that they had tried five or more diets in their lifetime (*USA TODAY*, June 21, 2011). Suppose that this result is true for the current population of adult women. A random sample of 13 adult women is selected. Use the binomial probabilities table (Table I of Appendix C) or technology to find the probability that the number of women in this sample of 13 who had tried five or more diets in their lifetime is

a. at most 7 **b.** 5 to 8 **c.** at least 7

Solution. Here, we have a sample of size $n = 13$ adult women, and the probability that one of them has tried five or more diets is $p = 50\% = 0.5$. Thus

a. P(at most 7) $=$ **binomcdf**$(13, 0.5, 7) \;=\; 0.7095$

b. P(5 to 8) $=$ **binomcdf**$(13, 0.5, 8) \;-\;$ **binomcdf**$(13, 0.5, 4) \;=\; 0.7332$

c. P(at least 7) $= 1 -$ **binomcdf**$(13, 0.5, 6) \;=\; 0.5000$

Chapter 6: Continuous Random Variables and the Normal Distribution

We consider here only the *normal distribution* for a continuous random variable x. A normally distributed random variable is referred to simply as a normal random variable. Its probabilities are determined by the *normal probability density function*, (or normal pdf, the familiar bell-shaped curve shown below), and the *normal cumulative distribution function* (or normal cdf, defining the *area* under the bell-shaped curve between any two points on the x-axis). The normal pdf is completely characterized by two *parameters*, its mean μ and its standard deviation σ, and is given in terms of them by the following exponential formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{15}$$

The normal pdf for a random variable x with mean $\mu = 30$ and standard deviation $\sigma = 5$ is illustrated in Figure 1. Note that the peak of a normal pdf always occurs at the mean value μ which, in Figure 1, is at $x = 30$.

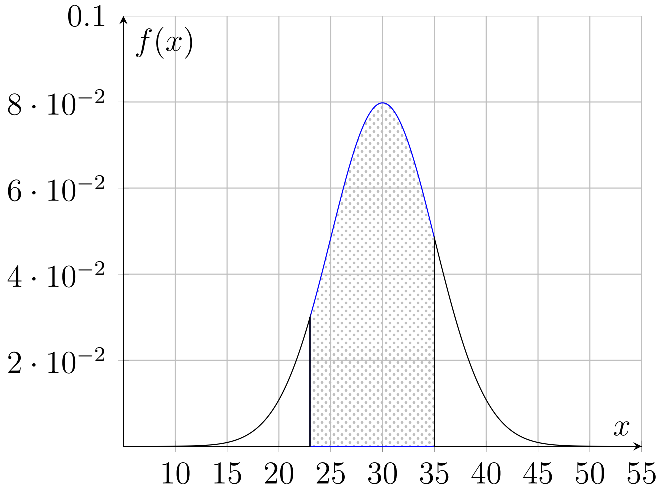

Figure 1: Normal pdf with $\mu = 30$, $\sigma = 5$

The *probability* that a normal random variable x lies in some interval between two values of x, say x_1 and x_2, with $x_1 < x_2$, is equal to the *area under the bell curve* between x_1 and x_2. In the above example, the probability that x lies between 23 and 35 is the shaded area under the bell curve between these two values of x. The probability is *calculated* using the normal cumulative distribution function (abbreviated as normal cdf) and the two given points. For two arbitrary points x_1 and x_2, this probability is written as $P(x_1 \leq x \leq x_2)$, so the shaded area in the example is equal to $P(23 \leq x \leq 35)$.

1. On the TI-84 press [2ND], then [VARS], and under the `DISTR` menu choose **normalcdf**. You must input the lower bound x_1 of the interval, the upper bound x_2 of the interval, the mean μ of the normal distribution, and the standard deviation

σ of the normal distribution. It should output a positive decimal number less than or equal to 1, which is the probability $P(x_1 \leq x \leq x_2)$ that x is in the interval $[\,x_1, x_2\,]$.

2. On the TI-83, after selecting **normalcdf** the function **normalcdf** will appear on the homescreen, where you must then input a comma–separated list consisting of the lower bound, the upper bound, the mean, and the standard deviation, that is, **normalcdf**(x_1, x_2, μ, σ). Press [ENTER] to output the probability.

The textbook focuses on the use of the *standard* normal distribution and its *table of values* in Table IV, Appendix C, page C19. A value in this Table represents, always, the area under the standard normal pdf to the *left* of a given value of z, defined as follows: an arbitrary normal random variable x with mean μ and standard deviation σ can be *transformed* to a *standard normal* random variable z defined by

$$z = \frac{x - \mu}{\sigma}. \tag{16}$$

The probability density function of this standard normal random variable has *special* values of the mean and standard deviation, namely, $\mu_s = 0$ and $\sigma_s = 1$ (the textbook does *not* use this s-subscript notation, but it seems appropriate to *distinguish* these special values associated with the *standard* normal pdf from those of an arbitrary normal pdf). Thus, the formula for the standard normal pdf is given by (15) after setting $\mu = 0$ and $\sigma = 1$:

$$f_s(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}}. \tag{17}$$

To use Table IV, the x-boundaries x_1 and x_2 must be transformed to z-boundaries, which are given by

$$z_1 = \frac{x_1 - \mu}{\sigma}, \qquad z_2 = \frac{x_2 - \mu}{\sigma}. \tag{18}$$

In the example with $\mu = 30$, $\sigma = 5$, and the interval endpoints $x_1 = 23$, $x_2 = 35$, the corresponding z-values are

$$z_1 = \frac{23 - 30}{5} = -1.4, \qquad z_2 = \frac{35 - 30}{5} = \quad 1.0\,.$$

A graph of the standard normal distribution for these parameters is illustrated in Figure 3 on the next page. A standard normal variable may, of course, also be used in calculator computations by substituting $\mu = 0$ and $\sigma = 1$ in the **normalcdf** program, and the computed z-values. In our example, above, where $\mu = 30$ and $\sigma = 5$, you will find that

$$P(23 \leq x \leq 35) = \textbf{normalcdf}(35, 42, 30, 5) = 0.7606,$$

rounded to four decimal places, while using the standard normal pdf parameters:

$$P(-1.4 \leq z \leq 1) = \textbf{normalcdf}(-1.4, 1, 0, 1) = 0.7606,$$

giving the same value. In general, it is always true that $P(x_1 \leq x \leq x_2) = P(z_1 \leq z \leq z_2)$, when the z-values are computed using (18).

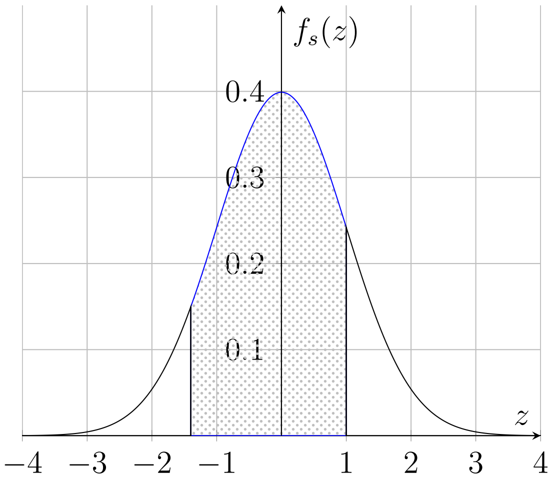

Figure 2: Standard Normal pdf.

You may also be asked to calculate $P(x \leq x_2)$, the probability that $x \leq x_2$ for some value of x_2. This is equivalent to asking for $P(-\infty < x \leq x_2)$, so the lower boundary x_1 is essentially "negative infinity" in this case. On the other hand, if you are asked to calculate $P(x \geq x_1)$, then this is equivalent to asking for $P(x_1 \leq x < \infty)$, so the upper boundary x_2 is essentially "positive infinity".

Infinity must be *approximated* on a calculator by some large finite number. It is probably safest to use 10^{99}, which you can input with the keystrokes [2ND], then [**,**] (the *comma* key to the right of the [x^2] key), then 99 (it will appear as 1E99 on your homescreen). So, for $x \leq x_2$, use **normalcdf**($-1\text{E}99, x_2, \mu, \sigma$) (don't forget the negative sign on 1E99, using the [(-)] key at the bottom right corner of the number keyboard, *not* the subtraction key [$-$]), and for $x \geq x_1$, use **normalcdf**($x_1, 1\text{E}99, \mu, \sigma$).

To find, for example, the probability that $x \leq x_2$ where $x_2 = 80$ and assuming that $\mu = 100$, $\sigma = 40$, first notice that the lower boundary is $-\infty$. We can then calculate either

$$P(x \leq 80) \;=\; \textbf{normalcdf}(-1\text{E}99, 80, 100, 40),$$

or instead compute $z_2 = (80 - 100)/40 = -0.5$, and calculate

$$P(z \leq -0.5) \;=\; \textbf{normalcdf}(-1\text{E}99, -0.5, 0, 1).$$

Both methods determine the *same* probability, that is, $P(x \leq 80) = P(z \leq -0.5) = 0.3085$ to four decimal places.

Another type of problem from this chapter asks you to find the value c of a normal random variable x corresponding to a *given probability* $P(x \leq c)$. This probability is just the *area*, say A_L, under the normal curve to the *left* of c.

To calculate the required value c of x on either the TI-84 or TI-83, press [2ND], then [VARS], and under the `DISTR` menu select **invNorm**.

1. For the TI-84, input the given probability (or area) A_L, the mean μ, and the standard deviation σ.

2. On the TI-83, **invNorm** appears on the homescreen, and you must input the comma–separated list A_L, μ, σ, which outputs

$$c \,=\, \textbf{invNorm}(A_L, \mu, \sigma), \qquad \text{where} \quad A_L \,=\, P(x \leq c). \tag{19}$$

If you are instead asked for the value c of x corresponding to a given probability $P(x \geq c)$, then this probability is the area to the *right* of c, say, A_R. Since the **invNorm** program requires as input an area to the *left* of c, you must now input $1 - A_R$ for the area in the **invNorm** program, that is,

$$c \,=\, \textbf{invNorm}(\,(1 - A_R),\, \mu,\, \sigma), \qquad \text{where} \quad A_R \,=\, P(x \geq c). \tag{20}$$

This follows from the fact that the total area under the normal curve is always 1, so $A_L + A_R = 1$, hence $A_L = 1 - A_R$.

Chapter 7: Sampling Distributions

The problems of this chapter involve *sampling from a population* that has a probability distribution for which either the mean μ *and* standard deviation σ are *known*, or the proportion p of some characteristic of the population is *known*.

- For each sample from a population having known μ and σ, we can calculate the mean $\overline{x}$ of that sample. This sample mean is a random variable, and the probability distribution of the sample means is referred to as the *sampling distribution* of $\overline{x}$. The mean of the sampling distribution is denoted by $\mu_{\overline{x}}$, and the standard deviation of the sampling distribution is denoted by $\sigma_{\overline{x}}$ (see Section 7.2 of [1]). The standard deviation $\sigma_{\overline{x}}$ of $\overline{x}$ is also called the *standard error of* $\overline{x}$. They are given in terms of the mean μ and standard deviation σ of the population from which they were sampled by

  $$\mu_{\overline{x}} = \mu\,, \qquad \text{and} \qquad \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}\,, \qquad \text{if} \quad \frac{n}{N} \leq .05\,, \tag{21}$$

  where n is the sample size, and N is the population size (there is a more complicated formula for $\sigma_{\overline{x}}$ if $n/N > .05$, but there are very few problems in the textbook that require it; see page 327 for details). Notice that if both the population standard deviation σ and sample standard deviation $\sigma_{\overline{x}}$ are given, then one can calculate the required sample size n by

  $$n = \left(\frac{\sigma}{\sigma_{\overline{x}}}\right)^2, \tag{22}$$

  which is the answer to a popular question on quizzes and exercises.

  If the population has a *normal* distribution with mean μ and standard deviation σ, then the *sampling* distribution is *also* normal, but with mean $\mu_{\overline{x}} = \mu$ and standard deviation $\sigma_{\overline{x}} = \sigma/\sqrt{n}$.

  If the population distribution has a known mean μ and standard deviation σ, but is *not* normally distributed, then the central limit theorem implies that for sample sizes $n \geq 30$, the sampling distribution will *again* be (approximately) normally distributed with mean $\mu_{\overline{x}} = \mu$ and standard deviation $\sigma_{\overline{x}} = \sigma/\sqrt{n}$.

  In summary, as long as $n/N \leq .05$ and the population is either normally distributed or, if not normally distributed, the sample size $n \geq 30$, we can assume that the sampling distribution is normal with mean $\mu_{\overline{x}} = \mu$ and standard deviation $\sigma_{\overline{x}} = \sigma/\sqrt{n}$. For a given sample we can then use, as in Chapter 6, the **normalcdf** function on the calculator to find the probability that the sample mean $\overline{x}$ lies between two given values, say $\overline{x}_1$ and $\overline{x}_2$, that is,

  $$P(\overline{x}_1 \leq \overline{x} \leq \overline{x}_2) = \textbf{normalcdf}\left(\overline{x}_1, \overline{x}_2, \mu, \frac{\sigma}{\sqrt{n}}\right).$$

  It is extremely important that the fourth argument be $\sigma/\sqrt{n}$, not simply σ, since we are considering a *sampling* distribution (that happens to be normal).

Again, the textbook focuses on using the standard normal distribution Table IV to calculate the probabilities. The sample mean $\overline{x}$ is transformed to a new *standard* normal random variable z defined by

$$z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{\sqrt{n}\,(\overline{x} - \mu)}{\sigma}, \tag{23}$$

where the second equality follows by substituting for $\sigma_{\overline{x}} = \sigma/\sqrt{n}$. This z is normally distributed with mean $\mu_s = 0$ and standard deviation $\sigma_s = 1$. The $\overline{x}$-interval $[\overline{x}_1, \overline{x}_2]$ transforms to a z-interval $[z_{\overline{x}_1}, z_{\overline{x}_2}]$, where

$$z_{\overline{x}_1} = \frac{\sqrt{n}\,(\overline{x}_1 - \mu)}{\sigma}, \qquad z_{\overline{x}_2} = \frac{\sqrt{n}\,(\overline{x}_2 - \mu)}{\sigma}.$$

These values are then used with Table IV to calculate the probability $P(\overline{x}_1 \leq \overline{x} \leq \overline{x}_2) = P(z_{\overline{x}_1} \leq z \leq z_{\overline{x}_2})$. This probability can also be calculated in terms of the parameters of the standard normal distribution using the same **normalcdf** program on the calculator:

$$P(\overline{x}_1 \leq \overline{x} \leq \overline{x_2}) = P(z_{\overline{x}_1} \leq z \leq z_{\overline{x}_2}) = \textbf{normalcdf}(z_{\overline{x}_1}, z_{\overline{x}_2}, 0, 1).$$

- The other type of problem discussed in this chapter is that of a population for which a *proportion* p share a *common characteristic*. If X is the number of elements of the population having this characteristic, and N is the population size, then the *population proportion* p with this characteristic is just the ratio (or fraction, or relative frequency) defined by

  $$p = \frac{X}{N}. \tag{24}$$

  If a sample of size n is taken from the population, and x elements in the sample have this characteristic, then the *sample proportion* with this characteristic is defined and denoted by

  $$\widehat{p} = \frac{x}{n}, \qquad \text{where } \widehat{p} \text{ is pronounced "p-hat"'}. \tag{25}$$

  The sample proportion is a random variable, and its probability distribution is called the sampling distribution of $\widehat{p}$. The mean $\mu_{\widehat{p}}$ of the sampling distribution of $\widehat{p}$ is always equal to the population proportion p:

  $$\mu_{\widehat{p}} = p.$$

  The standard deviation $\sigma_{\widehat{p}}$ of the sampling distribution of $\widehat{p}$ is given by the formula

  $$\sigma_{\widehat{p}} = \sqrt{\frac{p\,q}{n}}, \qquad \text{when} \quad \frac{n}{N} \leq .05\,,$$

  where $q = 1 - p$. See page 345 of the textbook for the calculation when $n/N \geq .05$.

In a given problem, you may be given p and $\widehat{p}$ directly, or you may be given the numbers X and x having a common characteristic, from which you can calculate the proportions: $p = X/N$ and $\widehat{p} = x/n$ (assuming the population and sample sizes have been given). Sometimes, *percentages* $p\,\%$ and $\widehat{p}\,\%$ having the common characteristic are given, and these should always be *converted to decimal fractions*, that is, *proportions*, by dividing them by 100: $p = p\,\%/100$, and $\widehat{p} = \widehat{p}\,\%/100$.

If the following two conditions are satisfied:

$$np > 5, \quad \text{and} \quad nq > 5, \quad \text{where} \quad q = 1 - p,$$

then the central limit theorem implies that the *sampling* distribution of $\widehat{p}$ is *approximately normal*. If these two conditions are met, we can compute the probability that the sample proportion lies between two values $\widehat{p}_1$ and $\widehat{p}_2$ again using the **normalcdf** program on the calculator:

$$P(\widehat{p}_1 \leq \widehat{p} \leq \widehat{p}_2) \;=\; \textbf{normalcdf}\left(\widehat{p}_1, \widehat{p}_2, p, \sqrt{\frac{pq}{n}}\right).$$

Be aware that the interval in question may be unbounded in either direction, so that you may be required to use $\widehat{p}_1 = -1E99$, or $\widehat{p}_2 = 1E99$. If desired, the standard normal distribution Table IV may be used after transforming $\widehat{p}$ to a standard normal random variable z defined by

$$z = \frac{\widehat{p} - \mu_{\widehat{p}}}{\sigma_{\widehat{p}}} = \frac{(\,\widehat{p} - p\,)}{\sqrt{\dfrac{p\,q}{n}}}.$$

Chapter 8: Estimation of the Mean and Proportion

We are considering in this chapter either a population random variable x, or a population proportion p. The random variable x has a population distribution whose population mean μ, however, is now an *unknown* value, and the proportion p of the population distribution of the proportion is also *unknown*. We are given a *single* sample set of data from one of these populations, and are interested in *estimating* (or approximating) the unknown values μ and p using this sample.

From the sample we can calculate the sample mean $\overline{x}$, or sample proportion $\widehat{p}$. The sampling distribution for $\overline{x}$ will have some mean $\mu_{\overline{x}}$, and the sampling distribution for $\widehat{p}$ will have some mean $\mu_{\widehat{p}}$. If we *knew* the population mean μ and population proportion p, we could conclude as in Chapter 7 that $\mu_{\overline{x}} = \mu$ and $\mu_{\widehat{p}} = p$. However, we do *not* know μ or p, so in this chapter we try to *estimate* values for them.

The simplest type of estimate we can make is called a *point estimate*. These are simply

$$\mu = \overline{x}, \qquad \text{and} \qquad p = \widehat{p}, \qquad \text{the point estimates of } \mu \text{ and } p.$$

A more conservative approach is to define a *confidence interval* that *contains*, with some specified reliability, either the population mean, or the population proportion. The reliability associated with the confidence interval is stated as a *percent* called the *confidence level*, denoted by $(1-\alpha)\,100\,\%$. In this definition, $(1-\alpha)$ (a decimal less than 1) is called the *confidence coefficient*, and α is called the *significance level*. The textbook does not set aside a special symbol for the confidence coefficient, but it is convenient to do so, and we will refer to it as simply $\mathcal{C}$, that is, $\mathcal{C} = 1 - \alpha$.

Though not explicitly mentioned in the textbook, α is always selected to be some small, *positive*, decimal value less than 0.5, that is, we always choose $0 < \alpha < .5$. The most common examples of confidence levels are 90 %, 95 %, 96 %, 97 %, 98 % , 99 %, and sometimes 99.5 %, corresponding to *confidence coefficients* of $\mathcal{C} = (1 - \alpha) = .900$, .950, .960, .970, .980, .990, and .995, the result of choosing *significance levels* of $\alpha =$ .100, .050, .040, .030, .020, .010, and .005, respectively.

We focus first on estimating μ, using the *sample* mean $\overline{x}$. To do so, we require first that *either* the population is normally distributed, *or* the sample size n is at least 30: $n \geq 30$. In these two cases, the *sampling distribution* is normal or approximately so, with mean $\mu_{\overline{x}} = \mu$. There are then *two possibilities* to consider: either the population standard deviation σ is known, or it is unknown (the population mean μ is of course *unknown*, as it is what we are trying to estimate).

1. σ known (z-values) If σ is *known*, then for any significance level $\alpha <$ .5 we can find a *positive* value of z (for which we use the *special symbol* $z_{\alpha/2}$), defined such that the area under the standard normal curve between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is $1-\alpha = \mathcal{C}$, the confidence coefficient.

The value $z_{\alpha/2}$ can be computed with either calculator using the **invNorm** function described in Chapter 6, and will correspond to one of the entries under the ***z* Value** column of Table 8.1, page 366, of the textbook. In this case, the area to the *left* of the *positive* value $z_{\alpha/2}$ is $1 - \dfrac{1-\mathcal{C}}{2} = \dfrac{1+\mathcal{C}}{2}$, so $z_{\alpha/2}$ can be calculated as

$$z_{\alpha/2} = \textbf{invNorm}\left(\left(\frac{1+\mathcal{C}}{2}\right), 0, 1\right).$$

When the standard deviation σ of a population is *known*, from which we then take a sample of size n and compute the sample mean $\overline{x}$, a $(1-\alpha)\,100\,\%$ *confidence interval* is *defined* by

$$\left[\overline{x} - \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}},\, \overline{x} + \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}\right] \tag{26}$$

This *confidence interval* can be calculated using either the TI-83 or TI-84 by

1. pressing [STAT],
2. scrolling over to `TESTS`,
3. selecting **ZInterval**, and pressing [ENTER].

Then, under **ZInterval**, at `Inpt`, scroll right and highlight **Stats** by pressing [ENTER] again.

Input the known population standard deviation σ, the sample mean $\overline{x}$, the sample size n, and the (decimal valued) confidence coefficient $(1-\alpha)$ (not the confidence level, which is a percent). At the bottom, highlight **Calculate**, then press [ENTER].

The top line of the output will contain the confidence interval in the form (x_L, x_U), where the numbers x_L and x_U are the lower and upper endpoints, respectively, of the confidence interval. You will also see the values for $\overline{x}$ and n that you entered as **Stats**.

The term being added and subtracted from $\overline{x}$ in the confidence interval (26) is called the *margin of error*, denoted by E:

$$E = \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}. \tag{27}$$

In order for the margin of error to have a *specified value* for a given α and σ, the sample size n must be chosen to be the nearest integer *greater* than

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2, \tag{28}$$

obtained by solving algebraically the previous equation for n.

2. σ **unknown (*t*-values)** If the standard deviation σ of the normal population distribution is *unknown*, then we begin by either being given, or calculating using the methods of Chapter 3, the *sample standard deviation* s (which is denoted by Sx in the output of the program **1-Var Stats** used in Chapter 3). Then $\overline{x}$ is a known random variable (with unknown mean $\mu_{\overline{x}} = \mu$, but known *sample* standard deviation $s_{\overline{x}} = s/\sqrt{n}$). Notice that $s_{\overline{x}}$ is NOT the same as the input variable Sx to your calculator program, which is s itself. A new random variable t defined by

$$t = \frac{\sqrt{n}\,(\overline{x} - \mu)}{s},$$

is used to determine the confidence interval in this case. However, t is *not* a *standard normal* random variable, but rather is determined by *Student's t distribution* (typically abbreviated to just t distribution). This distribution is also symmetric about a mean $\mu_t = 0$, but its standard deviation is given by $\sigma_t = \sqrt{\dfrac{df}{df - 2}}$ where $df = n - 1$ is called the number of *degrees of freedom*, and n is again the sample size.

When σ is *unknown*, the $(1 - \alpha)\,100\,\%$ confidence interval for the population mean μ is given by

$$\left[\overline{x} - \frac{t_{\alpha/2}\, s}{\sqrt{n}},\ \overline{x} + \frac{t_{\alpha/2}\, s}{\sqrt{n}}\right].$$

This confidence interval is calculated using either the TI-83 or TI-84 by

1. pressing [STAT],
2. scrolling over to `TESTS`,
3. selecting **TInterval**, and pressing [ENTER].

Then, under **TInterval**, at `Inpt`, scroll over and highlight **Stats** by pressing [ENTER] again.

Input the given, or calculated, *sample* standard deviation s for Sx, the sample mean $\overline{x}$, the sample size n, and the (decimal valued) confidence coefficient $\mathcal{C}$ for the C-Level. At the bottom, highlight **Calculate**, then press [ENTER].

The top line of the output will contain the confidence interval in the form (t_L, t_U), where the numbers t_L and t_U are the lower and upper endpoints, respectively, of the confidence interval. You will also see the values for $\overline{x}$, s, and n that you entered as **Stats**.

But how do we find $t_{\alpha/2}$, which defines the confidence interval? For a given significance level α, the *positive* value $t_{\alpha/2}$ of t is given by

$$P(t > t_{\alpha/2}) = \alpha/2.$$

On the TI-84, the value $t_{\alpha/2}$ can be calculated using the **invT** program obtained by pressing [2ND], then [VARS], and selecting it from the `DISTR` menu. Since the area to the *left* of $t_{\alpha/2}$ is $\dfrac{1+\mathcal{C}}{2}$, input $\dfrac{1+\mathcal{C}}{2}$ where **invT** asks for area, and then the number of degrees of freedom, df (which may have to be calculated from the sample size using $df = n - 1$, when the sample size n is given in the problem). When you press [ENTER], the output is the desired value $t_{\alpha/2}$, that is,

$$t_{\alpha/2} = \textbf{invT}\left(\left(\frac{1+\mathcal{C}}{2} \right), df \right).$$

Unfortunately, **invT** is *not included as an application* on the TI-83 calculator! However, based on the instructions given at

`http://www.ehow.com/how_7483505_download-invt-function-ti-calculator.html`

an inverse t distribution program (named **INVTW**) was written for the TI-83, which can be downloaded from

`http://mikewilkes-irsc.weebly.com/probability-and-statistics.html`

After downloading to your computer, you must send it to your TI-83. Links to directions for making this transfer are given at the above website.

The margin of error for a t confidence interval is

$$E = \frac{t_{\alpha/2}\, s}{\sqrt{n}}.$$

For a *specified* margin of error E, with given s and α, one might think that the required sample size would be

$$n = \left(\frac{t_{\alpha/2}\, s}{E} \right)^2.$$

However, in order to compute the t-value $t_{\alpha/2}$, one must know the number of degrees of freedom, which in turn requires the sample size n. Since n is unknown, this formula cannot be used. We instead use equation (28) with the population standard deviation σ replaced by the sample standard deviation s, that is,

$$n = \left(\frac{z_{\alpha/2}\, s}{E} \right)^2. \tag{29}$$

3. **Proportion Intervals** Finally, we consider a confidence interval estimate for an unknown population proportion p using a known sample proportion $\widehat{p}$. As in Chapter 7, the mean $\mu_{\widehat{p}}$ of the sampling distribution of $\widehat{p}$ is equal to the population proportion p:

$$\mu_{\widehat{p}} = p,$$

and the standard deviation $\sigma_{\widehat{p}}$ of the sampling distribution is given by

$$\sigma_{\widehat{p}} = \sqrt{\frac{p\,q}{n}}, \qquad \text{when} \quad \frac{n}{N} \leq .05\,,$$

where $q = 1 - p$. However, since the population proportion p is in this case *unknown*, we must *estimate it* by the *sample proportion* $\widehat{p}$ in the calculation *of the mean and standard deviation* $\mu_{\widehat{p}}$ and $\sigma_{\widehat{p}}$, obtaining for the sampling distribution statistics:

$$\mu_{\widehat{p}} = \widehat{p}, \qquad \text{and} \qquad \sigma_{\widehat{p}} = \sqrt{\frac{\widehat{p}\,\widehat{q}}{n}}\,, \qquad \text{when} \quad \frac{n}{N} \le .05\,, \tag{30}$$

where $\widehat{q} = 1 - \widehat{p}$. The sampling distribution can be considered to be *normal* as long as the following two conditions are met:

$$n\widehat{p} > 5, \quad \text{and} \quad n\widehat{q} > 5.$$

As in the previous development of a confidence interval for μ, we can find the positive value $z_{\alpha/2}$ of the standard normal random variable defined such that

$$P(z > z_{\alpha/2}) = \alpha/2 = \frac{1+\mathcal{C}}{2},$$

using the **invNorm** program on the calculator:

$$z_{\alpha/2} = \textbf{invNorm}\left(\left(\frac{1+\mathcal{C}}{2}\right), 0, 1\right).$$

The $(1-\alpha)\,100\,\%$ confidence interval for the population proportion p is given by

$$\left[\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}\,\widehat{q}}{n}}\,,\, \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}\,\widehat{q}}{n}}\right]. \tag{31}$$

This confidence interval is calculated using either the TI-83 or TI-84 by

1. pressing [STAT],
2. scrolling over to `TESTS`,
3. selecting **1-PropZInt**, and pressing [ENTER].

Under **1-PropZInt**, the value entered for `x` should be $n * \widehat{p}$, *rounded to the nearest integer* (rounding to obtain an *integer* is important; if you forget to do this, the calculator will remind you with an ERR:DOMAIN error response). Enter the sample size for n, and the confidence level $\mathcal{C}$ for C-Level. At the bottom, highlight **Calculate**, then press [ENTER]. The top line of the output will contain the confidence interval in the form $(\widehat{p}_L, \widehat{p}_U)$, where the numbers $\widehat{p}_L$ and $\widehat{p}_U$ are the lower and upper endpoint sample proportions (decimal values less than or equal to 1), respectively, of the confidence interval. You will also see the values for $\widehat{p}$ and n that you used in calculating x for your input.

The margin of error for a p confidence interval is

$$E = z_{\alpha/2}\sqrt{\frac{\widehat{p}\,\widehat{q}}{n}}.$$

For a *specified* margin of error E and confidence level (from which $z_{\alpha/2}$ can be determined) this can be solved for the sample size:

$$n = \frac{\widehat{p}\,\widehat{q}\,(z_{\alpha/2})^2}{E^2}.$$

Chapter 9: Hypothesis Tests About the Mean and Proportion

This chapter involves hypothesis testing of a *null hypothesis* H_0 about either

1. a *population mean* μ under one of two conditions: (a) the population variance σ is *known*, or (b) the population variance σ is *unknown*, or

2. a *population proportion* p, where we have *large samples* satisfying the inequalities $np > 5$, and $nq > 5$, where $q = 1 - p$, and n is the sample size.

For hypothesis tests of a population mean μ, the possible *null hypotheses* are either (i) $H_0 : \mu = \mu_0$, (ii) $H_0 : \mu \leq \mu_0$, or (iii) $H_0 : \mu \geq \mu_0$. For a population proportion the possible *null hypotheses* are (i) $H_0 : p = p_0$, (ii) $H_0 : p \leq p_0$, or (iii) $H_0 : p \geq p_0$. Notice that null hypotheses *always* have an *equality sign* included as *part* of the hypothesis (and often include *only* an equality sign).

In the case of a population mean, there are three possible *alternative hypotheses*, either (i) $H_1 : \mu \neq \mu_0$ (equivalent to either $\mu < \mu_0$ or $\mu > \mu_0$), (ii) $H_1 : \mu > \mu_0$, or (iii) $H_1 : \mu < \mu_0$. For a population proportion the possible *alternative hypotheses* would be (i) $H_1 : p \neq p_0$ (equivalent to either $p < p_0$ or $p > p_0$), (ii) $H_1 : p > p_0$, or (iii) $H_1 : p < p_0$. Alternative hypotheses *never* include an equality sign.

A null hypothesis is *rejected* whenever its corresponding *alternative* hypothesis is *true*. If the alternative hypothesis is $H_1 : \mu \neq \mu_0$ or $H_1 : p \neq p_0$, the rejection region will be *two-tailed*. If instead the alternative hypothesis is either $H_1 : \mu > \mu_0$ or $H_1 : p > p_0$, the rejection region is *one-tailed* and lies to the *right* of the non-rejection region (in short, it is *right-tailed*), while if $H_1 : \mu < \mu_0$ or $H_1 : p < p_0$, the rejection region is again *one-tailed* but lies to the *left* of the non-rejection region (in short, it is *left-tailed*).

The *rejection regions* in each case are *determined* by the value of the significance level α, introduced earlier in Chapter 8. This significance level α is the probability of *rejecting* H_0, when it is in fact *true*:

$$\alpha \,=\, P(H_0 \text{ is rejected} \,|\, H_0 \text{ is true}).$$

The significance level α is also called *the probability of a Type I error*.

Critical Values

Certain **critical values** of z or t, *determined by* α, are associated with each of the three types of alternative hypotheses:

1. If the rejection region is two-tailed, and σ is **known**, then α determines two critical values, a positive one, $z_{\alpha/2}$, defined by

   $$P(z \,>\, z_{\alpha/2}) = \alpha/2\,, \tag{32}$$

   and a second critical value that is the negative of this value. Since the area to the left of the *positive* value $z_{\alpha/2}$ is $1 - \alpha/2$, we can calculate it using either the invNorm algorithm: $z_{\alpha/2} = \textbf{invNorm}(\,(1-\alpha/2),\,0,\,1)$, or by table lookup. The two values $\pm\, z_{\alpha/2}$ are called the *critical values of z* for *two-tailed* tests of either the population mean when σ is known, **or** of the *population proportion* when doing proportion problems.

If the rejection region is two-tailed, but σ is **unknown**, then α determines a positive value $t_{\alpha/2}$ such that

$$P(t > t_{\alpha/2}) = \alpha/2\,, \tag{33}$$

which is calculated on the TI-84 using $t_{\alpha/2} = \textbf{invT}(\,(1-\alpha/2),\,df\,)$ (see Chapter 8 for alternatives if you are using a TI-83). It can also be found by table lookup in Table V, pages C-21 and C-22 of the textbook (but be careful here as you must use the actual value of $\alpha/2$, rather than $(1-\alpha/2)$ in the table lookup, since *this* Table, unlike Table IV, gives the area to the *right* of a given t). The values $\pm t_{\alpha/2}$ are the *critical values of t* for *two-tailed* tests of the population mean when σ is unknown.

2. If the rejection region is right-tailed, and σ is **known**, then α determines a *different* positive value z_α using α, <u>not</u> $\alpha/2$, such that

$$P(z > z_\alpha) = \alpha\,. \tag{34}$$

This is calculated using $z_\alpha = \textbf{invNorm}(\,(1-\alpha),\,0,\,1)$. This value z_α is called the *critical value of z* for *right-tailed* tests of either the population mean when σ is known, **or** of the *population proportion* in a test about proportions.

If the rejection region is right-tailed, but σ is **unknown**, then α determines a different positive value t_α such that

$$P(t > t_\alpha) = \alpha\,, \tag{35}$$

which is again calculated on the TI-84 using $t_\alpha = \textbf{invT}(\,(1-\alpha),\,df\,)$. This t_α is the *critical value of t* for *right-tailed* tests of the population mean when σ is unknown.

3. Finally, if the rejection region is left-tailed, and σ is **known**, then α determines a *negative* value z_α, such that

$$P(z < z_\alpha) = \alpha\,. \tag{36}$$

In this case α is the area to the *left* of z_α, so we must use α in calculating $z_\alpha = \textbf{invNorm}(\,\alpha,\,0,\,1)$. This value z_α is called the *critical value of z* for *left-tailed* tests of either the population mean when σ is known, **or** of the *population proportion.*

If the rejection region is left-tailed, but σ is **unknown**, then α determines a *negative* value t_α such that

$$P(t < t_\alpha) = \alpha\,, \tag{37}$$

which can be calculated on the TI-84 using $t_\alpha = \textbf{invT}(\,\alpha,\,df\,)$. This t_α is the *critical value of t* for *left-tailed* tests of the population mean when σ is unknown. If you use the t-Tables here, you must change the sign on the value of t obtained from the Table (it is negative for a left-tail test).

Observed Values, or Test Statistics

The values μ_0 and p_0 of μ and p, claimed in the statements of the *null hypotheses*, are used to define random variables called the **observed values**, or **test statistics**, of each test.

For the null hypothesis population mean μ_0 when σ is **known**, the observed value is the standard normal random variable z_0 defined by

$$z_0 \;=\; \frac{\sqrt{n}\,(\,\overline{x}\;-\;\mu_0)}{\sigma}\,. \tag{38}$$

When σ is **unknown**, the observed value is a t-distributed random variable t_0 defined by

$$t_0 \;=\; \frac{\sqrt{n}\,(\,\overline{x}\;-\;\mu_0)}{s}\,, \tag{39}$$

where s is the sample standard deviation.

For the null hypothesis population proportion p_0, the observed value is another standard normal random variable z_{p_0}, defined by

$$z_{p_0} \;=\; \frac{(\widehat{p} - p_0)}{\sqrt{\dfrac{p_0\,q_0}{n}}}\,, \tag{40}$$

where $q_0 = 1 - p_0$ under the radical sign.

p–Values

With each of the *observed values* defined above is associated a *probability*, called its p–value, which is used in a p–value test to determine whether or not to reject the null hypothesis. After computing the appropriate *observed* value from one of equations (38)–(40), we have

1. For a two-tailed test:

   - The p–value for a null hypothesis involving μ_0 is defined, when σ is *known*, by

     $$p_{z_0} \;=\; 2\,P(z\;>\;|z_0|) \;=\; 2\,P(z\;<\;-|z_0|)\,. \tag{41}$$

     The probabilities on the right-hand sides can be calculated using either **normalcdf**$(|z_0|, 1\text{E}99, 0, 1)$, or **normalcdf**$(-1\text{E}99, -|z_0|, 0, 1)$. Either probability *must be multiplied by* 2 to obtain the p–value corresponding to the observed value (test statistic) z_0.

   - When σ is *unknown*, the p–value is

     $$p_{t_0} \;=\; 2\,P(t\;>\;|t_0|) \;=\; 2\,P(t\;<\;-|t_0|)\,. \tag{42}$$

     This can be calculated using either **tcdf**$(|t_0|, 1\text{E}99, df)$, or **tcdf**$(-1\text{E}99, -|t_0|, df)$. Be advised, **tcdf** is the t–cumulative distribution function, which we have *not used prior to this example*; find it by pressing [2ND], then [VARS], and searching under the `DISTR` menu.

- For a null hypothesis involving the population proportion p_0, the p–value is

$$p_{p_0} = 2\,P(z > |z_{p_0}|) = 2\,P(z < -|z_{p_0}|), \tag{43}$$

which can be calculated using either **normalcdf**$(|z_{p_0}|, 1\text{E}99, 0, 1)$, or **normal cdf**$(-1\text{E}99, -|z_{p_0}|, 0, 1)$.

2. For a one-tailed test:

- The p–value for a null hypothesis involving μ_0 is defined, when σ is *known*, by

$$p_{z_0} = \begin{cases} P(z > |z_0|) & \text{for a right-tailed test,} \\ P(z < -|z_0|) & \text{for a left-tailed test.} \end{cases} \tag{44}$$

It can be calculated using either **normalcdf**$(|z_0|, 1\text{E}99, 0, 1)$, or **normalcdf**$(-1\text{E}99, -|z_0|, 0, 1)$.

- When σ is *unknown*, the p–value is

$$p_{t_0} = \begin{cases} P(t > |t_0|) & \text{for a right-tailed test,} \\ P(t < -|t_0|) & \text{for a left-tailed test.} \end{cases} \tag{45}$$

This can be calculated using either **tcdf**$(|t_0|, 1\text{E}99, df)$, or **tcdf**$(-1\text{E}99, -|t_0|, df)$, using the t–cumulative distribution function.

- For a null hypothesis involving the population proportion p_0, the p–value is

$$p_{p_0} = \begin{cases} P(z > |z_{p_0}|) & \text{for a right-tailed test,} \\ P(z < -|z_{p_0}|) & \text{for a left-tailed test.} \end{cases} \tag{46}$$

which can be calculated using either **normalcdf**$(|z_{p_0}|, 1\text{E}99, 0, 1)$, or **normal cdf**$(-1\text{E}99, -|z_{p_0}|, 0, 1)$.

Tests for Rejection of the Null Hypothesis

- ***p*–value Tests**

The null hypothesis is **rejected**, under the various alternative hypotheses, according to the following *p–value* comparisons with the *significance level* α:

1. For tests of a population mean, if H_1 is any test (either one- or two-tailed) at significance level α, then for *known* σ, H_0 is **rejected** if $p_{z_0} < \alpha$, while for *unknown* σ, H_0 is **rejected** if $p_{t_0} < \alpha$.
2. For tests of a population proportion, if H_1 is any test (either one- or two-tailed) at significance level α, then H_0 is **rejected** if $p_{p0} < \alpha$.

- **Critical Value Tests**

  The null hypothesis is also **rejected**, under the various alternative hypotheses, according to the following *critical value* comparisons with the calculated *observed values*:

  Two-tailed:

  1. For tests of a population mean, if H_1 is a *two-tailed* test with critical value $z_{\alpha/2}$ for known σ, or $t_{\alpha/2}$ for unknown σ, then H_0 is rejected if $|z_0| > z_{\alpha/2}$ for known σ, or $|t_0| > t_{\alpha/2}$ for unknown σ. Note that for two-tailed tests, it is the *absolute value* of the observed value that is compared to the critical value (which has positive values in each of these cases).
  2. For tests of a population proportion, if H_1 is a *two-tailed* test with critical value $z_{\alpha/2}$, then H_0 is rejected if $|z_{p0}| > z_{\alpha/2}$.

  Right-tailed:

  1. For tests of a population mean, if H_1 is a *right-tailed* test with critical value z_α for known σ, or t_α for unknown σ, then H_0 is rejected if $z_0 > z_\alpha$ for known σ, or $t_0 > t_\alpha$ for unknown σ.
  2. For tests of a population proportion, if H_1 is a *right-tailed* test with critical value z_α, then H_0 is rejected if $z_{p0} > z_\alpha$.

  Left-tailed:

  1. For tests of a population mean, if H_1 is a *left-tailed* test with critical value z_α for known σ, or t_α for unknown σ, then H_0 is rejected if $z_0 < z_\alpha$ for known σ, or $t_0 < t_\alpha$ for unknown σ.
  2. For tests of a population proportion, if H_1 is a *left-tailed* test with critical value z_α, then H_0 is rejected if $z_{p0} < z_\alpha$.

Calculator Applications

The TI-83 and TI-84 will calculate *p–values* and *observed values* for the various types of alternative hypotheses, as follows:

1. For hypotheses about a claimed population mean μ_0 when σ is *known*, go to [STAT], then under TESTS select **Z-Test**. If you know the sample mean $\overline{x}$ and sample size n, scroll over and highlight **Stats**. Enter the hypothesized mean μ_0, the known σ, the sample mean $\overline{x}$, and the sample size n. If it is a two-tailed test select $\mu :\neq \mu_0$, if left-tailed select $< \mu_0$, and if right-tailed select $> \mu_0$. Highlight **Calculate** and press [ENTER]. In the output, z is the *observed value* z_0 calculated using equation (38), and p is the p–value (44) corresponding to it.

2. For hypotheses about a claimed population mean μ_0 when σ is *unknown*, go to [STAT], then under TESTS select **T-Test**. If you know the sample mean $\overline{x}$, sample standard deviation s, and sample size n, scroll over and highlight **Stats**. Enter the hypothesized mean μ_0, the sample mean $\overline{x}$, the known s, and the sample size n. If it is a two-tailed test select $\mu :\neq \mu_0$, if left-tailed select $< \mu_0$, and if right-tailed select $> \mu_0$. Highlight **Calculate** and press [ENTER]. In the output, t is the *observed value* t_0 calculated using equation (39), and p is the p–value (45) corresponding to it.

3. For hypotheses about a claimed population proportion p_0, go to [STAT], then under TESTS select **1-PropZTest**. Enter the hypothesized proportion p_0, the number x of samples *having the desired characteristic*, and the sample size n. If you are given in the problem the sample proportion $\widehat{p}$, rather than x directly, you must compute $x = n * \widehat{p}$, and input this value *rounded to the nearest integer*. If it is a two-tailed test select $\mu :\neq \mu_0$, if left-tailed select $< \mu_0$, and if right-tailed select $> \mu_0$. Highlight **Calculate** and press [ENTER]. In the output, z is the *observed value* z_{p0} calculated using equation (40), and p is the p–value (46) corresponding to it.

Caution!

The calculator does *not* perform the actual *hypothesis test* for you! The three calculator applications just described involved *no information* about the significance level α *defining* the rejection region. You, the student, must complete the hypothesis test for a given value of α by either (i) comparing *the p-value obtained from the calculator* to α . . . reject H_0 if p-value $< \alpha$, otherwise fail to reject, or (ii) comparing the calculated *observed value* to one of the *critical values* defined by equations (32)–(37), each of which is determined by either $\alpha/2$ (two-tailed tests), or α (one-tailed tests) . . . reject H_0 if the observed value lies in a rejection region defined by either $\alpha/2$, or α, respectively, fail to reject otherwise. These tests were discussed earlier under the headings p–value Tests and Critical Value Tests.

A Type of Problem Requiring the Use of $t-$Tables

There is a type of problem, illustrated by Exercise 9.62, page 432 of [1], that asks a question that cannot be answered by using a graphing calculator. For completeness, the problem is restated here:

Problem “A soft-drink manufacturer claims that its 12-ounce cans do not contain, on average, more than 30 calories. A random sample of 64 cans of this soft drink, which were checked for calories, contained a mean of 32 calories with a standard deviation of 3 calories. Does the sample information support the alternative hypothesis that the manufacturer’s claim is false? Use a significance level of 5%. Find the **range** for the p-value for this test. What will your conclusions be using this p-value and $\alpha = .05$?”

Solution The first question one must answer is: what is the null hypothesis of the problem? Let μ_0 be the average or mean number of calories in a 12-ounce can of this soft drink. The manufacturer claims that μ_0 is not, "on average, more than 30 calories." If μ_0 is not more than 30 calories, then it must be either less than or equal to 30 calories, so this claim is the null hypothesis of the problem, that is, $H_0 : \mu_0 \leq 30$ calories. The alternative hypothesis is the logical negation of H_0, hence $H_a : \mu_0 > 30$ calories, which requires a *right-tailed test*. Since a population standard deviation σ is not given in the problem statement, we must use the t distribution to test the hypothesis.

The first question asks if the sample information *supports* the *alternative hypothesis* H_a, using a significance level of 5%, that is, for $\alpha = .05$. This is the same as asking if the sample information requires us to *reject* the *null hypothesis* H_0 at this significance level! The sample mean is $\overline{x} = 32$, the sample standard deviation is $s = 3$, the sample size is $n = 64$, and the number of degrees of freedom is $df = n - 1 = 63$. The *critical* value of t, that is, $t_\alpha = t_{.05}$, for a right-tailed test and significance level $\alpha = .05$, is the value $t_{.05}$ defined by equation (35):

$$P(t > t_{.05}) = .05\,.$$

Using the TI-84 calculator, it is given by $t_{.05} =$ **invT**$(\,(1 - .05),\, 63\,) =$ **invT**$(\,.95,\, 63\,)$ $= 1.6694$. The *observed* value, or *test statistic*, t_0, is given by equation (39):

$$t_0 = \frac{\sqrt{n}\,(\,\overline{x} - \mu_0)}{s} = \frac{\sqrt{64}\,(\,32 - 30)}{3} = 5.3333\,.$$

Since the observed value $t_0 = 5.3333$ is (much) greater than the critical value $t_\alpha = 1.6694$ (t_0 lies deep in the rejection region to the right of t_α), we would by the critical value test reject the null hypothesis at this significance level, that is, the sample information *does* support the alternative hypothesis that the manufacturer's claim is false. Note that the critical value $t_{.05} = 1.669$ could also be found in the t distribution Table V of [1] at the intersection of $df = 63$ in the first *column* of the Table, and the area .05 in the right tail (the significance level) from the first *row* of the Table.

Once the observed value t_0 has been determined, the appropriate p-value p_{t_0} is given by equation (45):

$$p_{5.3333} = P(t > 5.3333).$$

It can be calculated using $p_{5.3333} =$ **tcdf**$(5.3333, 1\text{E}99, 63) = 6.953 \times 10^{-7}$. Since p_{t_0} is much less than the significance level α, that is, $6.953 \times 10^{-7} < .05$, we would again reject the null hypothesis, this time by the p-value test.

We should note that the observed value (test statistic) t_0 and its corresponding p-value, could have been determined by pressing the [STAT] key on the calculator, then under TESTS selecting **T-Test**. Scrolling right to highlight **Stats**, and entering the hypothesized mean μ_0, the sample mean $\overline{x}$, the sample standard deviation s, and the sample size n, the resulting output would include t, the *observed value* t_0 calculated using equation (39), and p, the p–value p_{t_0} from equation (45) corresponding to t_0. They are the same values determined in the above discussion.

We have now shown, using the calculator, that either test gives the same result, namely, to reject the null hypothesis and conclude that the manufacturer's claim is false.

However, the problem also asks you to find the *range* for the p-value for this test. This seems, at first sight, to be a strange request. After all, there is only one p-value, p_{t_0}, corresponding to a given observed value t_0, and we determined it using the calculator to be the quite small value of 6.953×10^{-7}.

A range of p-values is an *artifact* of attempting to use Table V of the textbook, rather than the calculator, to determine the p-value. Remember that this is the value of the area in the right tail, from the first row of Table V, corresponding to $t_0 = 5.3333$ and $df = 63$. We see that the largest value of t in the row for $df = 63$, however, is $t = 3.225$, for which the p-value is $p_{3.225} = .001$. The test statistic $t_0 = 5.3333$ lies somewhere to the *right* of this t-value if the table could be extended to include it, and the corresponding p-values in the first row *decrease* as we move to the right. As $t \to \infty$, the p-value approaches 0, so the p-value we seek for $t_0 = 5.3333$ lies somewhere *between* .001 and 0, that is, the p-value lies in the *range* $0 < p_{t_0} < .001$, concluding the solution to the problem using Table V. Since the significance level $\alpha = .05$ is greater than the largest p-value .001 of this range, we would again reject the null hypothesis based on this range of p-values.

Short Notes on Hypothesis Testing

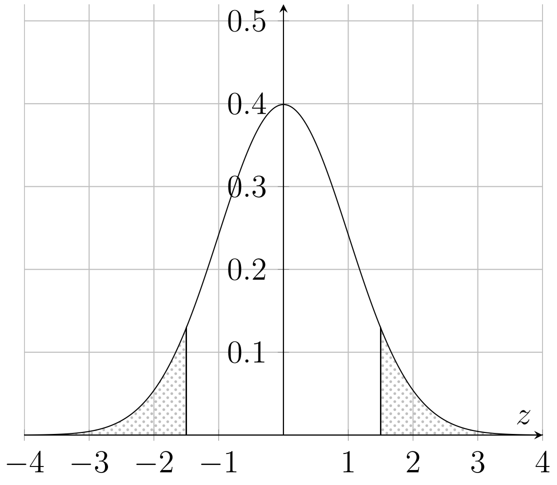

Figure 3: Standard Normal pdf.

Chapter 13: Simple Linear Regression

A simple linear regression model is the only topic to be discussed from this chapter. It is a probabilistic model that hypothesizes a linear relation between measurements of observations from either a population, or a sample from a population, of values of a dependent variable y, called the observed or actual values of y, and measured values of some independent variable x related to the measurements of y. Thus, two sets of data are involved, each containing N points if the data are from populations, or n points if they are from a *sample* of a population.

A *scatter plot* of the observed values y versus the independent variable values x can be obtained by first entering the x-values in list L1, and the y-values in list L2. Press the [Y=] key at the far upper left corner of the keyboard to display the equation editor. Press [CLEAR] at the `Y1=` tab to delete any function expression that may have been previously entered. Scroll up to `PLOT1` and press [ENTER] to turn off `PLOT1` (it will be highlighted after this step). Press [2ND] then [Y=], to go to the STAT PLOTS tab. Press [ENTER] to select `1:`, then highlight [ON] to turn [PLOT1] on for plotting data points. Scroll down to `TYPE`, and select the first option to create a scatter plot. At `XLIST` enter L1, and at `YLIST` enter L2. Scroll down to select a `Mark`, such as +, to indicate the symbol to be used for a point of the plot. Finally, press the [ZOOM] key, scroll down to `ZoomStat`, and press [ENTER] to obtain the scatter plot of your data.

NOTE. In order to return use of the `Y1=` tab in the equation editor to the graphing of a *function*, you must scroll up to `PLOT1` and press [ENTER] to remove the highlight, turning the plot back on. Failure to do so can be a great source of frustration when you wish later to graph, say, a polynomial function for your College Algebra course.

For linear regression, the *model* is an assumed straight-line relationship between estimated, or predicted, values of y, denoted by $\widehat{y}$, and the independent variable x of the data set:

$$\widehat{y} = a + bx\,. \tag{47}$$

We say that this line gives *the regression of variable y on variable x* . The $\widehat{y}$–intercept a and slope b are determined by finding the minimum value of the *error sum of squares SSE*, defined to be the sum of the squared (vertical) deviations of the observed y-values from the corresponding estimated values:

$$SSE = \sum (y - \widehat{y})^2. \tag{48}$$

The "least-squares" method for determining the linear regression coefficients a and b typically relies on concepts from *calculus*, see, for example, [7, Section 13.9], beyond the math prerequisites for the Elementary Statistics course. However, the coefficients can also be derived *using only algebra*, and in a concluding subsection we give the details of one such method due to Ehrenberg [8] that should be understandable by students who have had a College Algebra course.

Formulas Using the Sums and Sums of Squares of Data from Two Data Sets

A simple linear regression model is a probabilistic model that hypothesizes a linear relation between measurements of observed values of some *dependent* variable y, and measurements of values of some *independent* variable x. Thus, two sets of data are involved, each containing N points if the data are from a population, or n points if they are from a *sample* from a population. The *model* is an assumed straight-line relationship between *estimated or predicted* values of y, denoted by $\widehat{y}$, and the independent variable x:

$$\widehat{y} = a + bx\,. \tag{49}$$

We say that this line gives *the regression of variable y on variable x* .

The goal in all of this is the calculation of a and b to be used in equation (49). For problems in which you are given the sums $\sum x$ and $\sum y$, and the sums of squares or products, $\sum x^2$, $\sum y^2$, and $\sum xy$, follow these steps: From the textbook [1, Section 13.2.2], the three "sums of squares" SS_{xx}, SS_{xy}, and SS_{yy} are defined and calculated as follows (they are actually sums of *products* of *deviations from the mean*):

$$SS_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}, \tag{50}$$

$$SS_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}, \tag{51}$$

$$SS_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n}. \tag{52}$$

where n is the number of sample data points in each set. Formulas (50)–(52) hold for *sample data*. If *population data* is given, replace n by N in every formula.

The slope coefficient b is given in terms of two of the sums of squares as

$$b = \frac{SS_{xy}}{SS_{xx}}, \tag{53}$$

while the intercept a is given in terms of b and the sums of the x and y values by

$$a = \frac{\sum y}{n} - \frac{b \cdot \sum x}{n}, \tag{54}$$

with b given by equation (68). Equations (53) and (54) are all that are necessary to determine the linear regression equation (49). Just substitute these values of a and b into equation (49).

These values of the sums of squares provide the following *minimum SSE* of value

$$SSE_{min} = SS_{yy}\bigl(1 - r^2\bigr). \tag{55}$$

where the *coefficient of determination*, r^2, is defined by

$$r^2 = \frac{(SS_{xy})^2}{SS_{xx} \cdot SS_{yy}}. \tag{56}$$

The minimum SSE defined by equation (70) is *zero* only when $r^2 = 1$, or $r = \pm 1$, where

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} \tag{57}$$

is referred to as the *simple linear correlation coefficient*. When $r = 1$, the data is said to have *perfect positive linear correlation*, and when $r = -1$ it has *perfect negative linear correlation*.

It is important to observe that SS_{xx} and SS_{yy} are always *positive* (they truly are sums of squared terms, by definition, unlike SS_{xy}). But both the correlation coefficient r, and the slope b of the regression line, are proportional to SS_{xy}, which can be *either positive or negative*. Hence, a regression line with *positive slope* b corresponds to a positive correlation coefficient r (and the data are said to be *positively correlated*), while a regression line with *negative slope* b corresponds to a negative correlation coefficient r (and the data are said to be *negatively correlated*).

Using a TI Graphing Calculator to Determine the Regression Line Statistics from Two Sets of Data

On either the TI-83 or TI-84, the coefficients a and b can be determined from the sample x-values in list L1, and the sample y-values in list L2. Press [STAT], select `CALC`, and scroll down, past the option **LinReg(ax + b)**, to **LinReg(a + bx)**. On the TI-84, input L1 for Xlist, and L2 for Ylist; leave `FreqList` empty, but for `Store RegEQ` press [VARS], scroll to `YVARS`, select `1:FUNCTION`, then select `1:Y1`. This returns you to **LinReg(a + bx)** with `Y1` entered at `Store RegEQ` (storing the computed regression line equation to be graphed later). For the TI-83, input the comma–separated lists L1,L2,Y1 (obtaining Y1 using the steps just outlined). The output displays the regression equation (47) in the form $y = a + bx$ on the first line, and the values for the slope b and $\widehat{y}$-intercept a. By pressing [GRAPH] the regression line and scatter plot are now plotted on the same graph for comparison.

As a major revision to this Section, I realized that the regression coefficients a and b, the coefficient of determination r^2, the linear correlation coefficient r, and the *standard deviation of errors* s_e for a linear regression problem can all be determined from two given sets of either sample or population data by first entering one set as x-values in list L1, and the second set as y-values in list L2 (you must determine from the context which set represents the variable x and which set represents y). The **LinRegTTest** program on either the TI-83 or TI-84 will make the calculations for you.

Press [STAT], select `TESTS`, and scroll down to the **LinRegTTest** program. Input the two lists, set `Freq:1`, and choose β & $\rho : \neq \mathbf{0}$. Leave `RegEQ:` empty. The first line of the output is the regression equation $y = a + bx$. It is followed by the number of degrees of freedom for the problem, `df =` , then the values for the $\widehat{y}$-intercept a and slope b of the regression equation. The standard deviation of errors is `s =` , followed by `r`^{2} `=` and `r =` . It also performs a t-test on the value of the slope β (population parameter estimated by b), and a hypothesis test for the null hypothesis $H_o : \beta = 0$ (equivalently, $\rho = 0$) against one of the alternatives:

- $H_1 : \beta \neq 0$ and $\rho \neq 0$ (β & $\rho \neq 0$)
- $H_1 : \beta < 0$ and $\rho < 0$ (β & $\rho < 0$)
- $H_1 : \beta > 0$ and $\rho > 0$ (β & $\rho > 0$)

Confidence Interval Estimate of B

This confidence interval can be calculated as before in Chapter 8 with some minor modifications. Using either the TI-83 or TI-84, press [STAT], scroll over to `TESTS`, select **TInterval**, and press [ENTER]. Then, under **TInterval**, at `Inpt`, scroll over and highlight **Stats** by pressing [ENTER] again.

In place of the sample mean $\overline{x}$, use your calculated value for the coefficient b, that is,

$$\overline{x} : \; b. \tag{58}$$

Instead of using for Sx the sample standard deviation s, divided by $\sqrt{n}$, you must now compute the *standard deviation of errors*, s_e, defined by

$$s_e \; = \; \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}}, \tag{59}$$

where SS_{xy} and SS_{yy} are defined in equations (64) and (65), and n is the number of degrees of freedom. Then, using this value for s_e, compute

$$s_b \; = \; \frac{s_e}{\sqrt{SS_{xx}}}, \tag{60}$$

where SS_{xx} is given by equation (63). To obtain the correct confidence interval, the final value to be entered in place of Sx must be

$$Sx : \; s_b\,\sqrt{n-1}. \tag{61}$$

The last modification is to replace the sample size n by *one less* than that value:

$$n : \; n - 1. \tag{62}$$

The (decimal valued) confidence coefficient "C-Level" is entered as usual.

To illustrate the use of **TInterval** in getting a confidence interval estimate for B, here is how it would work for Example 13-4 of the textbook. You are given $n = 7$, $b = .2525$, $SS_{xx} = 1772.8571$, and $s_e = 1.5939$. We first compute $s_b = s_e/\sqrt{SS_{xx}}$, which gives us $s_b = 1.5939/\sqrt{1772.8571} = 0.037855$. Then, under **TInterval**, with `Inpt`: **Stats** highlighted, enter

$$\begin{aligned} \overline{x}&: \quad 0.2525 \\ Sx&: \quad 0.037855\,\sqrt{6} \\ n&: \quad 6 \\ \text{C-Level}&: \quad 0.95 \end{aligned}$$

The top line of the output will contain the confidence interval in the form (t_L, t_U), where the numbers t_L and t_U are the lower and upper endpoints, respectively, of the confidence interval.

Appendix: Algebraic Derivation of the Regression Coefficients

We begin with the error sum of squares SSE defined by equation (48), using definition (47) of the linear regression equation to replace $\widehat{y}$:

$$SSE = \sum (y - \widehat{y})^2 = \sum \big[\, y - \big(\, a + bx \,\big)\big]^2 .$$

To be clear, we should write SSE as $SSE(a, b)$, to indicate that it is a function of the regression line coefficients a and b, but for convenience these arguments will be omitted in what follows. The trick leading to an algebraic solution for a and b is the following. In the bracketed expression, we *add and subtract* the sample mean $\overline{y}$ of the y data values, and b times the sample mean $\overline{x}$ of the x data values, that is, $b\overline{x}$ (so the net effect is simply to add *zero* inside the bracket, leaving the *value* of SSE unchanged), to write SSE as

$$SSE = \sum \Big(y + \underbrace{\overline{y} - \overline{y}} + \underbrace{b\overline{x} - b\overline{x}} - a - bx \Big)^2 .$$

The terms can then be regrouped to obtain

$$SSE = \sum \Big\{ (y - \overline{y}) - b(x - \overline{x}) - \big[\, a - (\overline{y} - b\overline{x}) \,\big] \Big\}^2 .$$

Squaring the sum of the three distinct terms inside the summation, noting that $(A + B + C)^2 = A^2 + B^2 + C^2 + 2AB + 2BC + 2CA$, yields

$$\begin{aligned} SSE = \sum \Big\{ & (y - \overline{y})^2 + b^2(x - \overline{x})^2 + \big[\, a - (\overline{y} - b\overline{x}) \,\big]^2 - 2b(x - \overline{x})(y - \overline{y}) \\ & + 2b\big[\, a - (\overline{y} - b\overline{x}) \,\big](x - \overline{x}) - 2\big[\, a - (\overline{y} - b\overline{x}) \,\big](y - \overline{y}) \Big\}. \end{aligned}$$

The quantities a, b, $\overline{x}$, and $\overline{y}$, hence also the combination $\big[\,a\,-\,(\overline{y}\,-\,b\overline{x})\,\big]$, are all *constants*, which can be taken outside the summation wherever they occur as factors, and this allows us to write the last equation for SSE as

$$\begin{aligned} SSE \,=& \sum (y-\overline{y})^2 \,+\, b^2 \sum (x-\overline{x})^2 \,+\, n\,\big[\,a\,-\,(\overline{y}\,-\,b\overline{x})\,\big]^2 \,-\, 2b \sum (x-\overline{x})(y-\overline{y}) \\ &+\, 2b\big[\,a\,-\,(\overline{y}\,-\,b\overline{x})\,\big] \sum (x-\overline{x}) \,-\, 2\big[\,a\,-\,(\overline{y}\,-\,b\overline{x})\,\big] \sum (y-\overline{y}). \end{aligned}$$

But the sums of the deviations from the means both vanish, that is, $\sum (x-\overline{x}) = 0$ and $\sum (y-\overline{y}) = 0$, so we are left with

$$SSE \,=\, n\,\big[\,a\,-\,(\overline{y}\,-\,b\overline{x})\,\big]^2 \,+\, b^2 \sum (x-\overline{x})^2 \,-\, 2b \sum (x-\overline{x})(y-\overline{y}) \,+\, \sum (y-\overline{y})^2,$$

after rearranging terms. In accordance with the textbook [1, Section 13.2.2], we introduce the three "sums of squares" (really, these are "sums of products of deviations from the mean") SS_{xx}, SS_{xy}, and SS_{yy} defined by

$$SS_{xx} \,=\, \sum (x-\overline{x})^2 \,=\, \sum x^2 \,-\, n\overline{x}^2 \,=\, \sum x^2 \,-\, \frac{\left(\sum x\right)^2}{n}, \tag{63}$$

$$SS_{xy} \,=\, \sum (x-\overline{x})(y-\overline{y}) \,=\, \sum xy \,-\, n\overline{x}\,\overline{y} \,=\, \sum xy \,-\, \frac{\left(\sum x\right)\left(\sum y\right)}{n}, \tag{64}$$

$$SS_{yy} \,=\, \sum (y-\overline{y})^2 \,=\, \sum y^2 \,-\, n\overline{y}^2 \,=\, \sum y^2 \,-\, \frac{\left(\sum y\right)^2}{n}, \tag{65}$$

and substitute these definitions for the summations in the last equation for SSE:

$$SSE \,=\, n\,\big[\,a\,-\,(\overline{y}\,-\,b\overline{x})\,\big]^2 \,+\, b^2\, SS_{xx} \,-\, 2b\, SS_{xy} \,+\, SS_{yy}.$$

Factoring SS_{xx} from the two terms involving b, and completing the square on b in the result, we obtain

$$\begin{aligned} SSE \,&=\, n\,\big[\,a\,-\,(\overline{y}\,-\,b\overline{x})\,\big]^2 \,+\, SS_{xx}\left(b^2 \,-\, 2b\,\frac{SS_{xy}}{SS_{xx}}\right) \,+\, SS_{yy} \\ &=\, n\,\big[\,a\,-\,(\overline{y}\,-\,b\overline{x})\,\big]^2 \,+\, SS_{xx}\left(b \,-\, \frac{SS_{xy}}{SS_{xx}}\right)^2 \,-\, \frac{SS_{xy}^2}{SS_{xx}} \,+\, SS_{yy}. \end{aligned}$$

The *coefficient of determination*, r^2, is defined by

$$r^2 \,=\, \frac{SS_{xy}^2}{SS_{xx}\cdot SS_{yy}}, \tag{66}$$

which allows us to replace the third term of the SSE depending on SS_{xy}^2, by $r^2\, SS_{yy}$, yielding after a little algebra:

$$SSE \,=\, n\,\big[\,a\,-\,(\overline{y}\,-\,b\overline{x})\,\big]^2 \,+\, SS_{xx}\left(b \,-\, \frac{SS_{xy}}{SS_{xx}}\right)^2 \,+\, SS_{yy}\big(1\,-\,r^2\,\big). \tag{67}$$

Since the constant, SS_{xx}, is a sum of squared values, it is necessarily positive, and certainly the sample size n is positive, hence the coefficients of both perfect squares defining the SSE are positive. It follows that any deviations of a and b from the values

$$b \, = \, \frac{SS_{xy}}{SS_{xx}}, \tag{68}$$

and

$$a \, = \, \overline{y} \, - \, \overline{x}\, b, \tag{69}$$

with b given by equation (68), can only *increase* the SSE. Therefore, these values provide the *minimum* SSE of value

$$SSE_{min} \, = \, SS_{yy}\big(1 \, - \, r^2\big). \tag{70}$$

This is zero only when $r^2 = 1$, or $r = \pm 1$ (*perfect* positive or negative linear correlation), where

$$r \, = \, \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}, \tag{71}$$

is the *simple linear correlation coefficient.*

It is perhaps worth noting that if the coefficient of determination r^2 is replaced by its definition (66), and in that definition we replace $b = SS_{xy}/SS_{xx}$, we find that

$$r^2 \, = \, \frac{SS_{xy}^2}{SS_{xx} \cdot SS_{yy}} \, = \, \frac{b\, SS_{xy}}{SS_{yy}},$$

hence the minimum SSE of equation (70) can be written as

$$SSE_{min} \, = \, SS_{yy}\big(1 \, - \, r^2\big) \, = \, SS_{yy}\big(1 \, - \, \frac{b\, SS_{xy}}{SS_{yy}}\big) \, = \, S_{yy} \, - \, b\, SS_{xy},$$

or, in terms of the standard deviation of errors s_e defined by (59):

$$SSE_{min} \, = \, (n-2)\, s_e^2\, . \tag{72}$$

Thus, the minimum error sum of squares SSE_{min} is proportional to the square of the standard deviation of errors (the variance of errors?).

References

[1] Mann, P. S., *Introductory Statistics*, John Wiley and Sons, New York, NY, 8th ed., 2013.

[2] Mann, P. S., *Introductory Statistics*, John Wiley and Sons, New York, NY, 7th ed., 2010.

[3] Cox, R. T., *The Algebra of Probable Inference*, Johns Hopkins Press, Baltimore, 1961.

[4] Jaynes, E. T., "A Book Review of *The Algebra of Probable Inference*," *The American Journal of Physics*, Vol. 31, 1963, pp. 66–67.

[5] Jaynes, E. T., *Probability Theory: The Logic of Science*, Cambridge University Press, London, 2003.

[6] Jeffreys, H., *Theory of Probability*, Oxford University Press, London, 3rd ed., 1961.

[7] Larson, R., Hostetler, R., and Edwards, B. H., *Calculus: Early Transcendental Functions*, Brooks/Cole, Cengage Learning, Belmont, CA, 4th ed., 2007.

[8] Ehrenberg, A. S. C., "Deriving the Least Squares Regression Equation," *The American Statistician*, Vol. 37, No. 3, 1983, pp. 232.